Evaluation of statistical methods used in the analysis of interrupted time series studies: a simulation study

Simon L. Turner[1], Andrew B. Forbes[1], Amalia Karahalios[1], Monica Taljaard[2,3] and Joanne E. McKenzie[1*]

Abstract

Background: Interrupted time series (ITS) studies are frequently used to evaluate the effects of population-level interventions or exposures. However, examination of the performance of statistical methods for this design has received relatively little attention.

Methods: We simulated continuous data to compare the performance of a set of statistical methods under a range of scenarios which included different level and slope changes, varying lengths of series and magnitudes of lag-1 autocorrelation. We also examined the performance of the Durbin-Watson (DW) test for detecting autocorrelation.

Results: All methods yielded unbiased estimates of the level and slope changes over all scenarios. The magnitude of autocorrelation was underestimated by all methods, however, restricted maximum likelihood (REML) yielded the least biased estimates. Underestimation of autocorrelation led to standard errors that were too small and coverage less than the nominal 95%. All methods performed better with longer time series, except for ordinary least squares (OLS) in the presence of autocorrelation and Newey-West for high values of autocorrelation. The DW test for the presence of autocorrelation performed poorly except for long series and large autocorrelation.

Conclusions: From the methods evaluated, OLS was the preferred method in series with fewer than 12 points, while in longer series, REML was preferred. The DW test should not be relied upon to detect autocorrelation, except when the series is long. Care is needed when interpreting results from all methods, given confidence intervals will generally be too narrow. Further research is required to develop better performing methods for ITS, especially for short series.

Keywords: Autocorrelation, Interrupted time series, Public health, Segmented regression, Statistical methods, Statistical simulation

Background

Interrupted time series (ITS) studies are frequently used to evaluate the impact of interventions or exposures that occur at a particular point in time [1–4]. Although randomised trials are the gold standard study design, randomisation may be infeasible in the case of policy evaluation or interventions that are implemented at a population level. Randomization also is not an option for retrospective evaluation of interventions or exposures such as natural disasters or pandemics. The use of an ITS design may be considered in these situations, as they are one of the strongest non-randomised experimental designs [2, 5–9].

In an ITS study, observations are collected at regular time points before and after an interruption, and often analysed in aggregate using a summary statistic (e.g.

* Correspondence: joanne.mckenzie@monash.edu
[1]School of Public Health and Preventive Medicine, Monash University, 533 St Kilda Road, Melbourne, Victoria, Australia
Full list of author information is available at the end of the article

Turner *et al. BMC Medical Research Methodology*   (2021) 21:181

Page 2 of 18

mean, proportion) within a time interval (e.g. weekly, monthly, or annually). A key feature of the design is that data from the pre-interruption interval can be used to estimate the underlying secular trend. When this trend is modelled correctly, it can be projected into the post-interruption interval, providing a counterfactual for what would have occurred in the absence of the interruption. From this counterfactual, a range of effect measures can be constructed that characterise the impact of the interruption. Two commonly used measures include the 'change in level' – which represents the sustained change immediately after the interruption, and the 'change in slope' – which represents the difference in trends before and after the interruption.

A key feature of time series data is that there is the potential for non-independence of consecutive data points (serial autocorrelation) [10]. In the presence of positive autocorrelation, statistical methods that do not account for this correlation will give spuriously small standard errors (SEs) [11]. Several statistical methods are available to account for autocorrelation, such as Prais-Winsten generalised least squares or the Newey-West correction to SEs, or to directly model the error, such as autoregressive integrated moving averages (ARIMA). Further, several methods are available for testing for the presence of autocorrelation, with the Durbin-Watson test for lag-1 autocorrelation being the most commonly used [4, 6]. While the performance of some of these methods has been examined for time series data [12, 13], their performance in the context of ITS studies has received relatively less attention [11, 14, 15].

In this study, we therefore aimed to examine the performance of a range of statistical methods for analysing ITS studies with a continuous outcome using segmented linear models. We restrict our evaluation to ITS designs where there is a single interruption, with an equal number of time points pre and post interruption, and with first order autoregressive errors. Furthermore, we only consider single series, excluding controlled and multi-site ITS. The structure of the paper is as follows: In Section 2, we begin by introducing a motivating example for this research. In Section 3, we describe the statistical model and estimation methods used in our simulation study. In Sections 4 and 5, we present the methods and results from the statistical simulation study. In Section 6, we return to our motivating example and demonstrate the impact of applying the methods outlined in Section 3. Finally, in Section 7 we present key findings and implications for practice.

Motivating example
Healthcare-associated infections (HAIs) are a common complication affecting patients in hospitals. *C. difficile (C difficile)* infection is an example of one such infection

that can cause serious gastrointestinal disease. As such, many countries require mandatory surveillance of *C difficile* infection rates in hospitals. When outbreaks of *C difficile* occur, the cleaning and disinfection regimes in hospitals are often changed in an attempt to reduce the infection rate. The routine collection of data in this context has led to many retrospective investigations of the effects of different interventions (e.g. novel disinfectants) to reduce *C difficile* infection using ITS data [16]. Hacek et al. [17] provides an example of such a study, where they examined the effect of terminal room cleaning with dilute bleach (Fig. 1) on the rate of patients (per 1000 patient days) with a positive test for *C difficile*. Data were aggregated at monthly intervals. The series was relatively short – a scenario which is not atypical of these studies – with 10 data points pre and 24 post the intervention [16]. In the context of HAIs, there is a tendency for consecutive data points to be more similar to each other, manifesting as 'clusters' of data points in time (Fig. 1). Fitting a segmented linear regression model to the data shows an apparent immediate decrease in the infection rate (level change), as well as a decrease in the trend (slope change). In the following section, we outline different statistical methods to estimate the model parameters and return to this example in Section 6, where we apply these methods and compare the results.

Methods
Interrupted time series (ITS): model and estimation methods
We begin by describing the statistical model and parameters used in our simulation study followed by a brief description of some common statistical estimation methods and the Durbin-Watson test for autocorrelation.

Statistical model
We use a segmented linear regression model with a single interruption, which can be written using the parameterisation proposed by Huitema and McKean [18, 19] as:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 D_t + \beta_3 [t - T_I] D_t + \varepsilon_t \qquad (1)$$

where Y_t represents the continuous outcome at time point t of N time points. D_t is an indicator variable that represents the post-interruption interval (i.e. $D_t = 1$ ($t \geq T_I$) where T_I represents the time of the interruption). The model parameters, β_0, β_1, β_2 and β_3 represent the intercept (e.g. baseline rate), slope in the pre-interruption interval, the change in level and the change in slope, respectively. The error term, ε_t, represents deviations from the fitted model, which are constructed as:

Fig. 1 Rate of *C. difficile* infections (per 1000 patient-days) pre and post bleach disinfection intervention per month
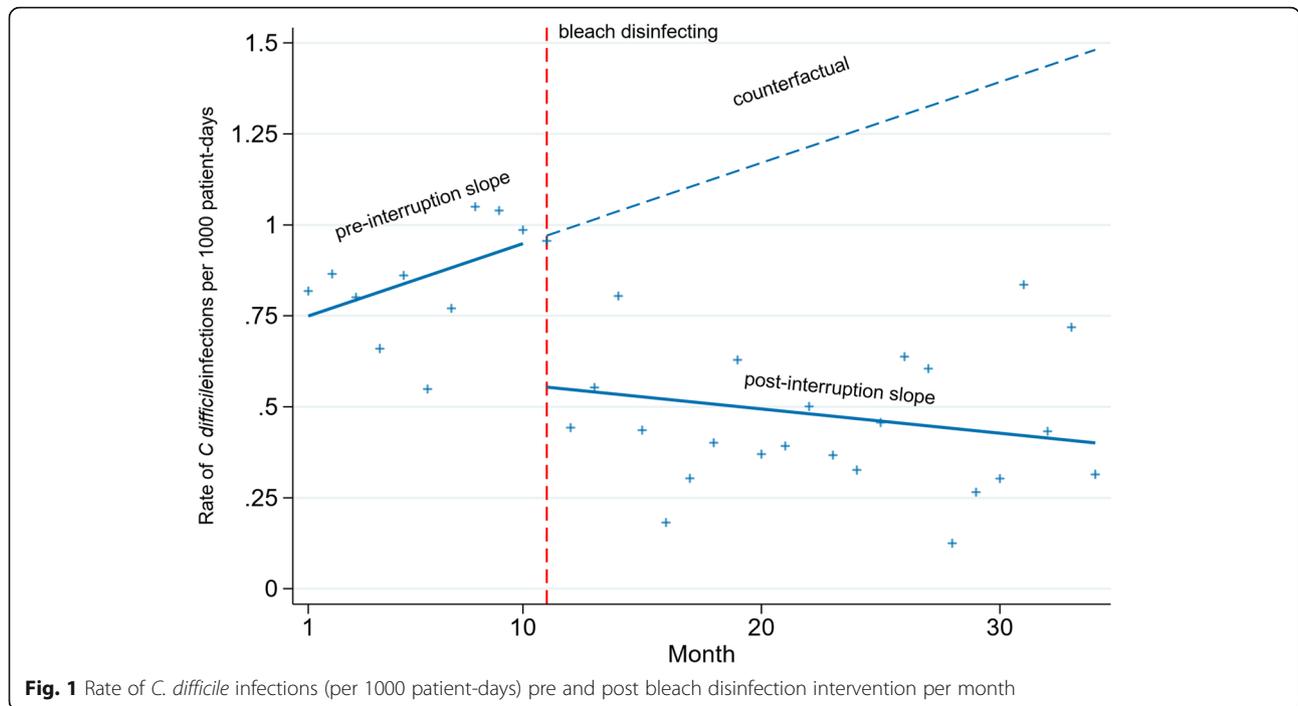
$$\varepsilon_t = \rho\varepsilon_{t-1} + w_t \tag{2}$$

where w_t represents "white noise" that is normally distributed $w_t \sim N(0, \sigma^2)$, and ρ is the lag-1 autocorrelation of the errors which can range from −1 to + 1. A lag-1 error means that the influence of errors on the current error is restricted to the value immediately prior. Longer lags are possible but in this paper we confine attention to lag-1 only (AR (1) errors).

Estimation methods

A range of statistical estimation methods are available for estimating the model parameters. These methods account for autocorrelation in different ways and are briefly described below. We focus on statistical methods that have been more commonly used (Ordinary Least Square (OLS), Generalised Least Squares (GLS), Newey-West (NW), Autoregressive Integrated Moving Average (ARIMA)) [2–4, 6]. In addition, we have included Restricted Maximum Likelihood (REML) (with and without the Satterthwaite adjustment), which although is not a method in common use, is included because of its potential for reduced bias in the estimation of the autocorrelation parameter, as has been discussed for general (non-interrupted) time series [20]. Further details and equations can be found in Additional file 1.

Ordinary least squares Estimates of the regression parameters and their variances from model [1] can be obtained from fitting a segmented linear regression model

using OLS (Additional file 1). In the presence of autocorrelation, the OLS estimators for the regression parameters are unbiased; however, the SEs will be incorrect [21].

Newey-West The NW estimator of the variance of the regression parameters estimated using OLS accommodates autocorrelation and heteroskedasticity of the error terms in the regression model (1) [22] (Additional file 1).

Generalised least squares Two common GLS methods for estimating the regression parameters and their variances are Cochrane-Orcutt (CO) and Prais-Winsten (PW). For both methods, a regression model is first fitted using OLS and an estimate of the autocorrelation is calculated from the residuals. This estimate is then used to transform the data and remove the autocorrelation from the errors, upon which the regression parameters are then estimated from the transformed data. If there is still some residual autocorrelation these steps are iterated until a criterion is met (e.g., the estimated value for autocorrelation has converged [23]). The CO method applies the transformation from the second observation onwards (t = 2, 3, … n). The PW method is a modification to the CO method in which a transformed value is used for the first observation (Additional file 1). The PW method is therefore likely to be more efficient in small series since it does not discard the first observation. The sampling properties of the estimators of the regression parameters are likely to be adversely affected when the

series length is small due to poor estimation of the autocorrelation.

Restricted maximum likelihood It is well known that maximum likelihood estimators of variance components are biased in small samples due to not accounting for the degrees of freedom (d.f.) used when estimating the fixed effect regression parameters [24]. Restricted maximum likelihood is a variant of maximum likelihood estimation and attempts to address the bias by separating the log-likelihood into two terms; one that involves the mean and variance parameters, and one which is only dependent on the variance parameters. By maximising the latter term first with the appropriate number of d.f., an estimate of the variance parameter can be obtained which can be used when maximising the former, thus correctly accounting for the d.f. [20, 25].

For small samples, there is greater uncertainty in the estimation of the SE of the regression parameters. To account for this uncertainty in making inferences about the regression parameters, the Satterthwaite adjustment can be used to adjust the t-distribution d.f. used in hypothesis testing and calculation of confidence limits [26].

ARIMA/ARMAX regression with autoregressive errors estimated using maximum likelihood A more flexible approach than the Prais-Winsten and Cochrane-Orcutt generalised least squares methods is called Autoregressive Moving Average eXogeneous modelling (ARMAX) [27]. Here we consider a simple case in which the exogenous variables are the functions of time to form a segmented regression model, and the errors are assumed to have an AR (1) structure. Parameters in this more general family of models are estimated by maximum likelihood, enabling the uncertainty in the autocorrelation estimate to be taken into account in the standard error of the regression coefficients, unlike PW or CO. This approach has been variously labelled in the literature, including use of the terminology 'maximum likelihood ARIMA' [14]. We therefore use the shorthand term "ARIMA" for consistency with previous literature, including in our companion paper [28]. Further details about the method can be found in Additional file 1, Paolella [27], Nelson [29] and Box et al. [30].

Durbin-Watson test for autocorrelation

The Durbin-Watson (DW) test is commonly used for detecting lag-1 autocorrelation in time series. Often, the test is used as part of a two-stage analysis strategy to determine whether to use a method that adjusts for autocorrelation or use OLS (which does not adjust for autocorrelation). The null hypothesis is that there is no autocorrelation ($H_0 : \rho = 0$) against the alternative that autocorrelation is present ($H_1 : \rho \neq 0$). The DW-statistic can range between zero and four, with values close to two indicating no autocorrelation. The DW-statistic is compared to critical values to determine whether there is evidence of autocorrelation, no autocorrelation, or the test is inconclusive. The critical values differ by series length, significance level and the d.f. in the regression model. Further details are available in Additional file 1, Kutner et al. [21] and Durbin and Watson [31].

Simulation study methods

We undertook a numerical simulation study, examining the performance of a set of statistical methods under a range of scenarios which included continuous data with different level and slope changes, varying lengths of series and magnitudes of lag-1 autocorrelation. Design parameter values were combined using a fully factorial design with 10,000 data sets generated per combination. A range of criteria were used to evaluate the performance of the statistical methods. We now describe the methods of the simulation study using the ADEMP (defining aims, data-generating mechanisms, estimands, methods and performance measures) structure [32].

Data generating mechanisms

We simulated continuous data from ITS studies by randomly sampling from a parametric model (Eq. 1), with a single interruption at the midpoint, and first order autoregressive errors (examples shown in Supplementary 1.1). We multiplied the first error term, ε_1, by $\sqrt{\frac{1}{1-\rho^2}}$ so that the variance of the error term was constant at all time points.

We created a range of simulation scenarios including different values of the model parameters and different numbers of data points per series (Table 1). These values were informed by our review of ITS studies [4], where we reanalysed available data sets to estimate level and slope changes (standardised by the residual standard deviation), and autocorrelation. We found a median standardised level change of 1.5 (inter-quartile range (IQR): 0.6 to 3.0), $n = 190$), median standardised slope change of 0.13 (IQR: 0.06 to 0.27, $n = 190$) and median autocorrelation 0.2 (IQR: 0 to 0.6, $n = 180$). We therefore constructed models with level changes (β_2) of 0, 0.5, 1 and 2, and slope changes (β_3) of 0 and 0.1. We did not examine negative level or slope changes since we did not expect this to influence the performance metrics. Lag-1 autocorrelation was varied between 0 and 0.8 in increments of 0.2 to cover the full range of autocorrelations observed in the ITS studies included in the review. The number of data points per series was varied from 6 to 100, equally divided before and after the interruption, informed by the number of data points observed in the ITS studies (median 48, IQR: 30 to 100, $n = 230$). The

Table 1 Simulation parameters

Parameter	Symbol	Parameter Values
Intercept	β_0	0
Pre-interruption slope	β_1	0
Level change	β_2	0, 0.5, 1, 2
Change in slope post-interruption	β_3	0, 0.1
Autocorrelation coefficient	ρ	0, 0.2, 0.4, 0.6, 0.8
Variance of white noise error component	σ^2	1
Number of data points		6, 8, 10, 12, 14, 16, 18, 20 24, 28, 32, 36, 40, 44, 48, 52, 56 60, 80, 100

increment between the number of data points per series varied; initially it was small (i.e. 2) so as to detect changes in the performance metrics that were expected to arise with smaller sample sizes and was increased to 4 and then 20.

All combinations of the factors in Table 1 were simulated, leading to 800 different simulation scenarios (Table 1, Fig. 2).

Estimands and other targets

The primary estimands of the simulation study are the parameters of the model, β_2 (level change) and β_3 (slope change) (Eq. 1). These were chosen as they are commonly reported effect measures [4, 6]. We also examined the autocorrelation coefficient, ρ, and the value of the Durbin Watson statistic.

Statistical methods to analyse ITS studies

Segmented linear regression models were fitted using the estimation methods described in Section 2.2. We evaluated estimation methods designed to estimate the model parameters under lag-1 autocorrelation (see Table 2 for details). For GLS, we restricted our

investigation to the PW method, because it was expected to have better performance than the CO method (on which PW is based) given the PW method utilises all data points. For REML with the Satterthwaite adjustment, we substituted d.f. of 2 when the computed d.f. were less than 2, to avoid overly conservative confidence limits and hypothesis tests. We also investigated the commonly used Durbin-Watson method for detecting autocorrelation at a significance level of 0.05 [31].

Table 2 summarises the methods and model variations used to adjust for autocorrelation. Details of the Stata code used for generating the simulated data and the analysis code can be found in the online repository figshare [33].

Performance measures

The performance of the methods was evaluated by examining bias, empirical SE, model-based SE, 95% confidence interval coverage and power (see Additional file 1 for formulae). Confidence intervals were calculated using the simsum package [34] with t-distribution critical values. For each simulation scenario, we used 10,000 repetitions in order to keep the Monte Carlo Standard Error (MCSE) below 0.5% for all potential values of
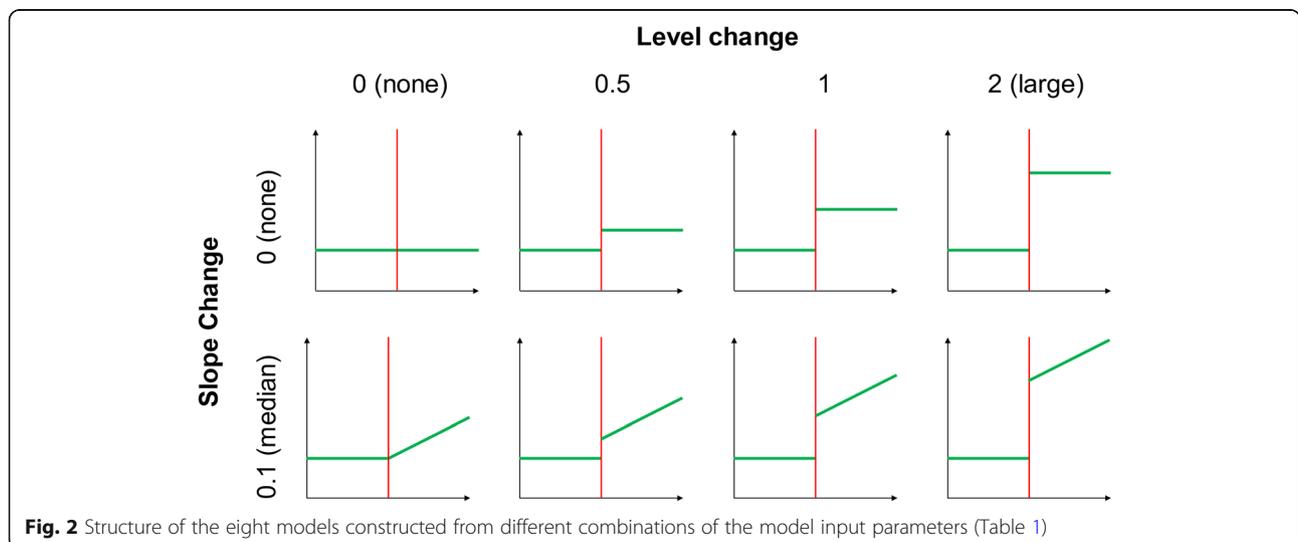
Fig. 2 Structure of the eight models constructed from different combinations of the model input parameters (Table 1)

Table 2 Statistical methods and adjustments for autocorrelation

Method	Autocorrelation adjustment
Ordinary Least Squares	None
	Newey-West SE adjustment (lag-1)
Generalised least squares	Prais-Winsten
Restricted maximum likelihood	Lag-1 autocorrelation
	Lag-1 autocorrelation with small sample Satterthwaite approximation
Autoregressive integrated moving average	Lag-1 autocorrelation (i.e. ARIMA(1,0,0))

coverage and type I error rate. Model non-convergence was recorded and tabulated.

Coding and execution

The statistical software Stata version 15 [35] was used for the generation of the simulated data. A random seed was set at the beginning of the process and the individual random state was recorded for each repetition of the simulated data sets. Each dataset was independently simulated, using consecutive randomly generated numbers from the starting seed. We used a "burn in" period between each dataset of 300 random number generations so that any lag effects specific to the computer-generated series had time to dissipate [11].

Prior to running the simulations, we undertook initial checks to confirm that the data generation mechanism was working as expected. This involved fitting series of length 100,000 to check the estimated β parameters matched the input parameters. A larger sample of 1000 datasets was then simulated and checked using summary statistics and graphs. When we were satisfied that the simulations were operating as expected, the full number of datasets were simulated.

Analysis of the simulated datasets

Analyses were performed using Stata version 15 [35]. A range of visual displays were constructed to compare the performance of the statistical methods. Frequency distributions were plotted to visualise the level- and slope-change estimates, autocorrelation coefficient estimates, and the results of the Durbin-Watson test for autocorrelation. Scatter plots were used to display the mean values for empirical and model-based SEs, coverage, power and autocorrelation coefficient estimates. Line plots were used to show confidence intervals for the level and slope change estimates. Results and summaries of the analyses were summarised (using the simsum package [34]) and graphed using Stata version 15 [35].

Results of the simulation study

Bias of level and slope change estimates

All methods yielded approximately unbiased estimates of level change and slope change across all simulation scenarios. Figure 3 presents level change estimates specific to the scenario of a level change of 2 and a slope change of 0.1 (Supplementary Fig. S2 shows slope change estimates), but the other 7 combinations of level and slope changes were virtually identical (Supplementary 1.3.1 for level change, Supplementary 1.3.2 for slope change). Note that the Satterthwaite and NW adjustments do not impact the parameter estimates of level or slope change, hence distributions of these parameter estimates are not shown in Figs. 3 and S2.

Standard errors of level and slope change estimates
Empirical standard errors

Figure 3 and Supplementary Fig. S2 visually indicate the precision of the estimators in terms of the spread of the distributions therein. To enable a direct quantitative assessment, we plotted the empirical SE of the level and slope changes for each method against selected series lengths and autocorrelation parameter sizes for a level change of 2 and slope change of 0.1 (Fig. 4 and Fig. 5). The size of the empirical SE of the level change was dependent on the underlying autocorrelation, length of the series and statistical method (Fig. 4). Of note, the estimates obtained from the ARIMA and PW models yield almost identical empirical SEs. For each magnitude of autocorrelation, the empirical SE decreased as the length of the time series increased, as would be expected. An exception to this occurred for the OLS estimator (and to a lesser extent ARIMA) which exhibited unusual behaviour for an autocorrelation of 0.8, with the SE initially increasing with an increasing number of points in the series, and then decreasing. Supplementary simulations were undertaken to examine the behaviour of the OLS estimator for surrounding correlations (0.7 and 0.9), which showed a similar pattern of increasing SEs with an increasing number of points (Supplementary 1.4). The relationship between autocorrelation and the empirical SE was modified by the length of series. For small series (< 10 data points), the empirical SE decreased with increasing autocorrelation, while for longer series (≥ 10 data points) this relationship was reversed, with SEs increasing with increasing autocorrelation.

The size of the empirical SE for slope change was dependent on the underlying autocorrelation and length

Fig. 3 Distributions of level change estimates calculated from four statistical methods, from top to bottom: autoregressive integrated moving average (ARIMA) (purple), ordinary least squares regression (OLS) (blue), Prais-Winsten (PW) (green) and restricted maximum likelihood (REML) (orange). The vertical axis shows the length of the time series. The five vertical columns display the results for different values of autocorrelation. The vertical black line represents the true parameter value (β_2). Each subset of four curves shows the distribution from a different analysis method for a given combination of time series length and autocorrelation. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other structures give similar results. The Satterthwaite adjustment to the REML method and the Newey-West adjustment to the OLS method do not impact the estimate of level or slope change, hence these parameter estimates are not shown

Fig. 4 Empirical standard error (SE) of the level change. The horizontal axis shows the length of the time series, the vertical axis shows the empirical SE. The five vertical columns display the results for different values of autocorrelation. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; PW, Prais-Winsten; REML, restricted maximum likelihood

Fig. 5 Empirical standard error (SE) of the slope change. The horizontal axis shows the length of the time series, the vertical axis shows the empirical SE. The five vertical columns display the results for different values of autocorrelation. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; PW, Prais-Winsten; REML, restricted maximum likelihood

of the series (Supplementary Fig. S2 and Fig. 5). The empirical SE decreased with increasing series length, but increased with increasing autocorrelation, as would be expected. In contrast to the level change, there were no important differences in the empirical SEs across the statistical methods, even when the autocorrelation was large. The observed patterns did not differ for any of the eight level and slope change combinations (Supplementary 1.3.3 for level change, Supplementary 1.3.4 for slope change).

Comparison between empirical and model-based standard errors

To enable appropriate confidence interval coverage and size of significance tests, the model-based SE needs to be similar to the empirical SE [32]. In this section we present the comparison between the empirical and model-based SEs; results for the model-based SEs alone can be found in S1.3.5 for level change and S1.3.6 for slope change.

For the level change parameter (β_2) estimated by OLS, the ratios of model-based to empirical SEs were close to one (indicating the empirical and model-based SEs were similar) for all series lengths when there was no underlying autocorrelation (Fig. 6). However, as autocorrelation increased, as expected, the OLS model-based SEs became increasingly smaller relative to the empirical SEs, indicating the model-based SEs are downwardly biased. The NW method performed only slightly better than the OLS (except when the autocorrelation was zero); however, the NW model-based SEs were still

downwardly biased across all scenarios, were worse than OLS for small series lengths, and only marginally better than OLS for large series lengths. Although the empirical SEs of the ARIMA and PW methods were similar, they had quite different model-based SEs. The PW model-based SEs were smaller than the empirical SEs for all magnitudes of autocorrelation, though the model-based SEs approached the empirical SEs with increasing series length. The ARIMA model-based SEs were larger than the empirical SEs for small series (fewer than 24 points) at small underlying values of autocorrelation ($\rho < 0.4$) and also for larger series (more than 24 points) at higher magnitudes of autocorrelation ($\rho > 0.4$). Aside from these scenarios, the ARIMA model-based SEs were approximately equal to the empirical SEs. The REML method behaved similarly to the PW method but, relatively, did not underestimate the SEs to the same extent. For small values of underlying autocorrelation ($\rho < 0.4$) and series greater than 30 points, the model-based SEs were similar to the empirical SEs.

For the slope change parameter (β_3), the ratios of model-based to empirical SEs followed similar patterns as for the level change parameter (β_2). For any given series length, as the magnitude of autocorrelation increased, model-based SEs became increasingly smaller compared with the empirical SEs for most statistical methods (Supplementary 1.5). Model-based and empirical SEs tended towards equivalence as series lengths increased, with the exception of OLS and NW at high values of autocorrelation ($\rho > 0.6$). For REML and ARIMA, the pattern of ratios of model-based to

Fig. 6 Scatter plots of the ratio of model-based standard error (SE) to the empirical SE for the level change parameter with different levels of autocorrelation and series length. The horizontal axis represents the number of points in the time series, the vertical axis shows the ratio of model-based to empirical SE. The five vertical columns display the results for different values of autocorrelation. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. The first two series lengths are not shown for the ARIMA method due to extreme values. The Satterthwaite adjustment to the REML does not impact the estimate of SE, hence details of this method are not shown. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood

empirical SEs for β_3 slightly differed compared with β_2. Specifically, the REML model-based SEs were smaller than the empirical SEs for small series, and then increased to be slightly larger as the number of points increased. For ARIMA, the model-based SEs were smaller than the empirical SEs for large underlying values of autocorrelation ($\rho \geq 0.6$) for small to moderate length series. The observed patterns did not differ for any of the eight level and slope change combinations (S 1.3.5 for level change, S 1.3.6 for slope change).

Confidence interval coverage

For all combinations of level change, slope change, number of time points and autocorrelation, most methods had coverage (percentage of 95% confidence intervals including the true parameter) that was less than the

nominal 95% level for both level and slope change (Fig. 7 for level change and Fig. 8 for slope change, both with a level change of 2 and slope change of 0.1, Supplementary 1.3.7 for level change and Supplementary 1.3.8 for slope change for other parameter combinations). The exceptions were OLS when there was no underlying autocorrelation, and REML with the Satterthwaite adjustment for moderate to large length series. In general, mean values of coverage decreased with increasing autocorrelation and increased with increasing series length. However, coverage of the OLS method decreased with increasing autocorrelation as well as with increasing series length (with the exception of the zero autocorrelation scenario). The NW method exhibited a similar pattern to OLS, but generally had better coverage (except for small autocorrelations), although coverage was often poor (under 90% for all but the longest series

Fig. 7 Coverage for the level change parameter. Each point is the proportion of the 10,000 simulations in which the 95% confidence interval included the true value of the parameter. The solid black line depicts the nominal 95% coverage level. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite

with low autocorrelation, $\rho < 0.4$). REML with the Satterthwaite small sample adjustment yielded coverage greater than the nominal 95% level when the number of data points was greater than 30 in the presence of autocorrelation. Confidence interval coverage patterns generally reflected those observed with the comparisons between the model-based and empirical SE.

Power

Coverage was less than the nominal 95% level in the majority of scenarios (except for the OLS model in the absence of autocorrelation and some scenarios involving the REML method with Satterthwaite adjustment). In scenarios where coverage is less than 95%, examining power is misleading. Due to there being only a very

Fig. 8 Coverage for the slope change parameter. Each point is the proportion of the 10,000 simulations in which the 95% confidence interval included the true value of the parameter. The solid black line depicts the nominal 95% coverage level. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite

small number of configurations in Fig. 7 and Supplementary 1.6 in which 95% coverage was achieved, we adopt a more liberal approach and consider configurations in which the coverage was at least 90%. As such, the results presented below should be viewed as approximate power only and will generally be lower than the value observed if coverage was at least 95%.

For scenarios with a level change of two, power was low for series with a small number of points, but predictably, increased as the number of points increased for all methods, except the REML method with Satterthwaite adjustment (Fig. 9). As the magnitude of autocorrelation increased its power decreased, to a point where it became lower than for other methods. This was due to the REML method with Satterthwaite adjustment having greater than 95% coverage in these situations and hence substantially lower than 5% Type I error rates. For smaller values of the level change parameter, predictably, power decreased (Supplementary 1.6.1). Similar patterns were observed for slope change (Supplementary 1.6.2).

Autocorrelation coefficient

Most of the statistical methods yield an estimate of the autocorrelation coefficient. All methods underestimated the autocorrelation for series with a small number of points (Fig. 10 and Fig. 11 show autocorrelation coefficient estimates for a simulation with parameter values of 2 for level change and 0.1 for slope change). However, underestimation was most pronounced for scenarios with small series and large underlying autocorrelation.

The REML method always yielded estimated autocorrelations closer to the true underlying autocorrelation compared with the other methods. The empirical SEs for autocorrelation generally decreased as the series length increased for all methods (except for small series with fewer than 20 points) (Supplementary 1.7). The observed patterns did not differ for any of the eight level and slope change combinations (Supplementary 1.3.9).

Durbin-Watson test for autocorrelation

The DW test for detecting autocorrelation performed poorly except for long data series and large underlying values of autocorrelation (Fig. 12). For series of moderate length (i.e. 48 points), with an underlying autocorrelation of 0.2, the DW test gave an "inconclusive" result in 30% of the simulations, incorrectly gave a value of no autocorrelation in 63% of the simulations, and only correctly identified that there was autocorrelation in 7% of the simulations. For shorter length series the percentage of simulations in which autocorrelation was correctly identified decreased (for a series length of 24 even at extreme magnitudes of autocorrelation (i.e. 0.8) positive autocorrelation was reported in only 26% of the simulations). For very short length series (fewer than 12 data points) the DW test gave an "inconclusive" result in over 75% of the simulations for all values of autocorrelation and always failed to identify that autocorrelation was present.
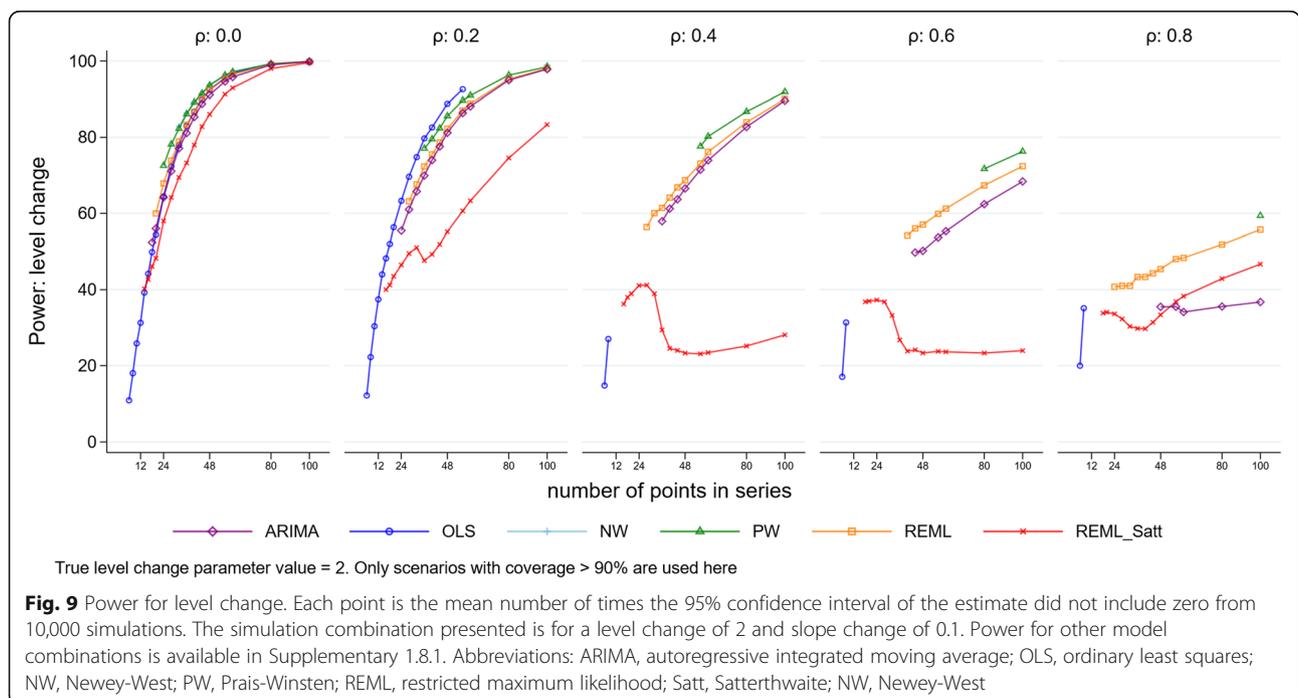
Fig. 9 Power for level change. Each point is the mean number of times the 95% confidence interval of the estimate did not include zero from 10,000 simulations. The simulation combination presented is for a level change of 2 and slope change of 0.1. Power for other model combinations is available in Supplementary 1.8.1. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite; NW, Newey-West
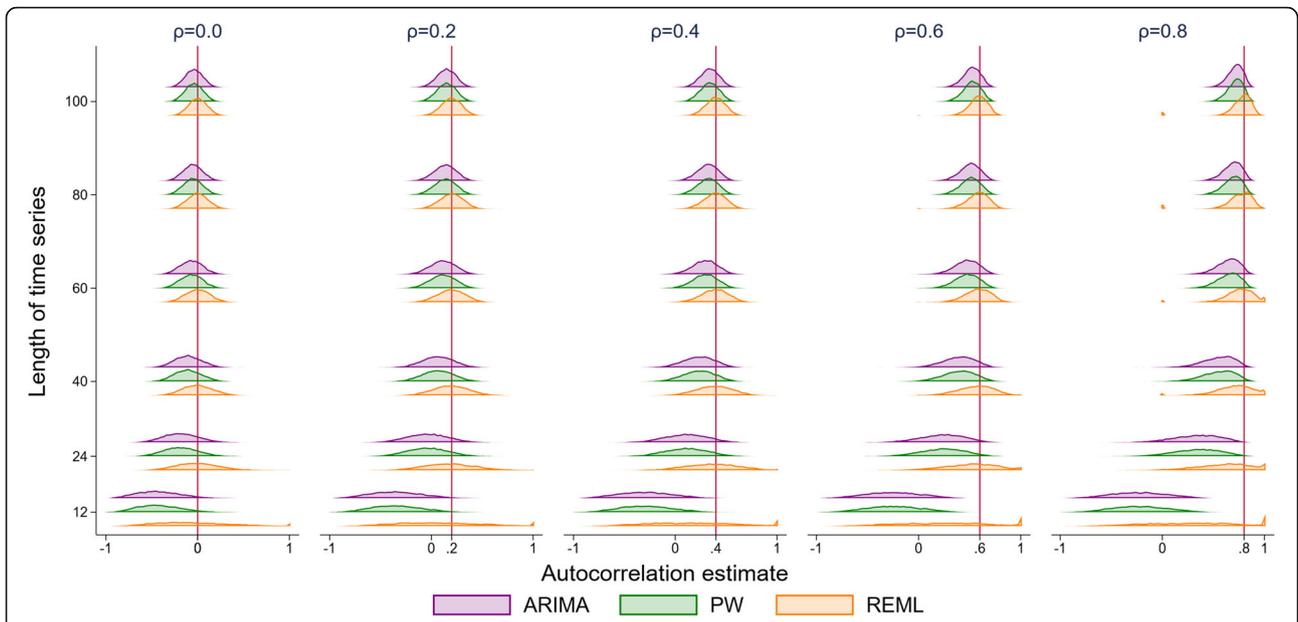
Fig. 10 Autocorrelation coefficient estimates. The horizontal axis shows the estimate of autocorrelation coefficient. The vertical axis shows the length of the time series. The five vertical columns display the results for different values of autocorrelation ranging from 0 to 0.8 (the value of autocorrelation is shown by a vertical red line). Each coloured curve shows the distribution of autocorrelation coefficient estimates from 10,000 simulations. Each subset of four curves shows the results from a different analysis method for a given combination of time series length and autocorrelation. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. From top to bottom the methods are autoregressive integrated moving average (ARIMA) (purple), Prais-Winsten (PW) (green) and restricted maximum likelihood (REML) (orange)
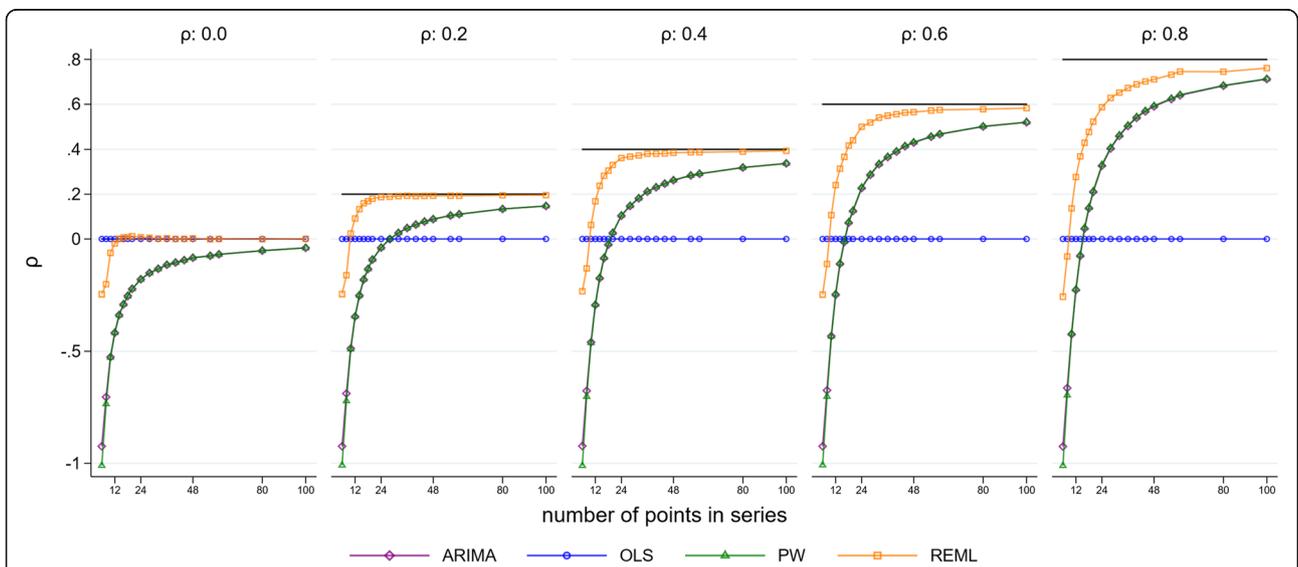
Fig. 11 Autocorrelation coefficient estimates. The horizontal axis shows the length of the time series. The vertical axis shows the mean estimate of the autocorrelation coefficient across 10,000 simulations. The five plots display the results for different values of autocorrelation ranging from 0 to 0.8 (the true value of autocorrelation is shown by a horizontal black line). Each coloured point shows the mean autocorrelation estimate for a given combination of true autocorrelation coefficient and number of points in the data series. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; PW, Prais-Winsten; REML, restricted maximum likelihood
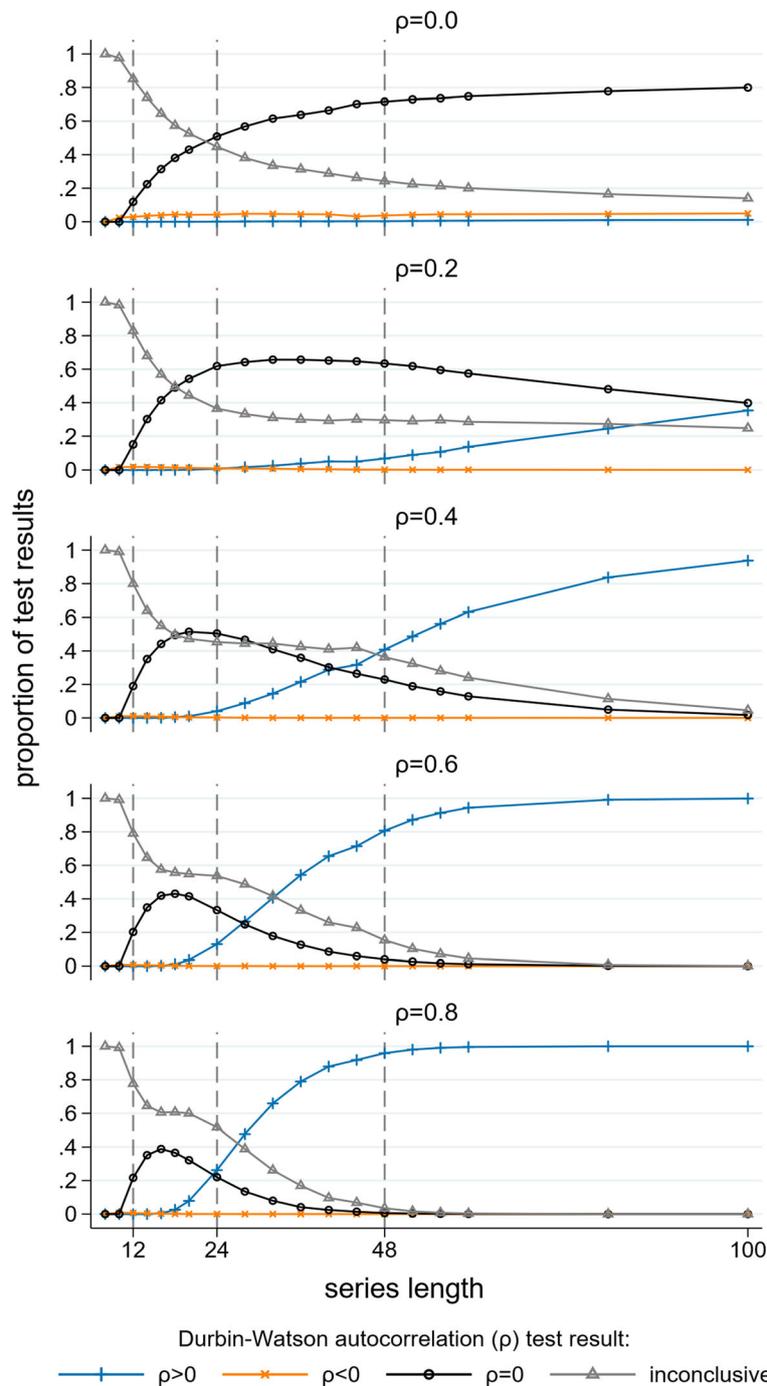
Fig. 12 Durbin-Watson tests for autocorrelation. For each combination of length of data series and true magnitude of autocorrelation the Durbin Watson test results from 10,000 simulated data sets are summarised. The horizontal axis is the length of the data series, the vertical axis is the proportion of results indicating: ρ > 0 (blue), ρ < 0, (orange) ρ = 0 (black) and an inconclusive test (grey). Each graph shows results for a different magnitude of autocorrelation. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results

Convergence of estimation methods

The number of the 10,000 simulations in which the estimation methods converged is presented in Supplementary 1.8. Most methods had no numerical convergence issues. The PW model failed to converge a small number of times (less than 7% of simulations) when there were only three data points pre- and post-interruption. The REML model regularly failed to converge (approximately

70% convergence) for short data series (fewer than 12 data points) at all values of autocorrelation, however convergence improved substantially as the number of points in the series increased. In addition, convergence issues for REML occurred more frequently for higher values of autocorrelation, unless the series length was large.

Analysis of motivating example

We re-analysed the ITS study (introduced in Section 2) using each of the statistical methods evaluated in the simulation study to estimate the effect of terminal room cleaning with dilute bleach on *C difficile* rates. Estimates of level and slope change (along with their confidence intervals and *p*-values) and autocorrelation are presented in Table 3. The point estimates for level and slope change are similar across methods, but notably, the width of the confidence intervals vary considerably. The confidence intervals are narrower for OLS, NW and PW, but wider for REML (with and without the Satterthwaite adjustment) and ARIMA. For level change, this led to corresponding *p*-values that ranged from 0.002 to 0.095; and for the slope change, p-values ranging from 0.069 to 0.531. Estimates of autocorrelation also varied, with REML yielding an estimate of 0.23, while ARIMA and PW yielded much lower estimates of 0.07. The DW statistic was 1.86, indicating no autocorrelation. Such differences in confidence interval width and p-values may impact on the interpretation of the results.

Discussion

Summary and discussion of key findings

Interrupted time series studies are commonly used to evaluate the effects of interventions or exposures. The results of our simulation study provide insight into how a set of statistical methods perform under a range of scenarios of continuous data which included different level and slope changes, varying lengths of series and magnitudes of lag-1 autocorrelation. We chose to examine

statistical methods that are commonly used in practice for interrupted time series studies [1–4, 6], and those performing well in the general, non-interrupted, time series literature [13, 20].

Not surprisingly, we found that the statistical methods all yielded unbiased estimates of both level and slope change for all values of model shape, length of series and autocorrelation. Confidence interval coverage, however, was generally below the nominal 95% level, except in particular circumstances for specific methods. The REML method with and without the Satterthwaite adjustment had improved confidence interval coverage compared with the other statistical methods, particularly for slope change. An exception to this was for very small series (fewer than 12 points), where the OLS method had better coverage than the other methods, even in the presence of large underlying autocorrelation. Coverage improved for most methods with increasing series length (with the exception of OLS and NW in some circumstances). REML with the Satterthwaite adjustment to the d.f. was the only method that yielded at least the nominal level of confidence interval coverage, however it was overly conservative in some scenarios, with a resultant reduction in power compared with other methods.

Autocorrelation was systematically underestimated by all statistical methods, with estimates of autocorrelation being particularly biased (and often negative) for small time series (fewer than 24 points). This underestimation of autocorrelation had a detrimental impact on the estimates of SE, which were too small, and in turn, this led to confidence interval coverage that was less than the nominal 95% level. This can be seen in Fig. 13 (level change) and Supplementary 1.9 (slope change), where a relationship between the magnitude of bias in the estimates of autocorrelation and confidence interval coverage is clearly evident. Ideally the confidence interval coverage should be at the nominal 95% level with no bias in autocorrelation (the intersection of the dashed lines in Fig. 13). For short time series, the severe underestimation of autocorrelation led to poorer confidence

Table 3 Level- and slope-change point estimates with 95% confidence intervals (CIs), p-values and estimate of magnitude of lag-1 autocorrelation ($\hat{\rho}_{est}$) from C difficile infection data using a range of statistical methods. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite

	Level change		Slope change		$\hat{\rho}_{est}$
	Estimate (CI)	p-value	Estimate (CI)	p-value	
ARIMA	−0.42 (−0.89,0.05)	0.079	−0.03 (−0.11,0.06)	0.531	0.07
OLS	−0.44 (−0.76,-0.13)	0.008	−0.03 (−0.07,0.02)	0.201	N/A
NW	−0.44 (−0.71,-0.17)	0.002	−0.03 (−0.06,0.00)	0.069	N/A
PW	−0.42 (−0.75,-0.09)	0.014	−0.03 (−0.08,0.02)	0.251	0.07
REML	−0.37 (−0.72,-0.01)	0.044	−0.02 (−0.08,0.03)	0.390	0.23
REML-Satt	−0.37 (−0.82,0.09)	0.095	−0.02 (−0.10,0.05)	0.437	N/A

Fig. 13 Bias in autocorrelation estimate versus coverage for level change. The horizontal axis shows the bias in the autocorrelation estimate. The vertical axis shows the percentage coverage. The horizontal dashed line indicates 95% coverage, the vertical dashed line indicates no bias in the estimate of autocorrelation. Each colour represents a different value of underlying autocorrelation, ranging from zero (purple) to 0.8 (red), with each value displayed in a circle at the smallest series length (six points). The arrows point from shortest to longest series length, with the small circles at the end of each line showing coverage at a series length of 100 data points. Each data point shows the mean value from 10,000 simulations for a given combination of autocorrelation coefficient and number of points in the series. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite

interval coverage than had autocorrelation been ignored, as is the case with OLS.

We included REML due to its potential to reduce bias in the variance parameters compared with maximum likelihood. Although the ARIMA model fitted in our simulations used maximum likelihood estimation, the model-based SEs were generally more similar to the empirical SEs for the ARIMA method compared with the REML method (where the model-based SEs were generally smaller than the empirical SEs). ARIMA confidence interval coverage was similar to REML for level change, though REML showed improved confidence interval coverage for slope change. Further, the REML method yielded less biased estimates of autocorrelation than the other methods, even for small series lengths.

The only method to yield overly conservative confidence intervals was the REML with SW adjustment to the t-distribution d.f.. When deciding whether to use the Satterthwaite adjustment, consideration therefore needs to be made between the trade-off in the risk of type I and type II errors. A further issue we identified with the Satterthwaite adjustment was that the adjusted d.f. were very small in some series, leading to nonsensible confidence intervals. To limit this issue we set a minimum value of 2 for the d.f., but other choices could be adopted.

The DW test is the most commonly used test to identify autocorrelation and is often used when series are short [4, 6]. Some authors use the test as part of a two-stage analysis strategy where they first test for autocorrelation, and depending on the result of the test, either use a method that attempts to adjust for autocorrelation or not. This type of two-stage approach is used in other contexts, such as testing for carryover in crossover trials. The findings of our

simulation study underscore why such two stage approaches fail and are discouraged; namely, due to their failure to detect the presence of a statistic when it exists (i.e., their high type II error rate). In our case, we found that for short series (fewer than 12 data points), the DW test failed to identify autocorrelation when it was present, and for moderate length series (i.e. 48 points), with an underlying autocorrelation of 0.2, autocorrelation was only detected in 7% of the simulations. Other tests for autocorrelation are available [15, 36–39], though they are not commonly used in practice [4, 6], and have been shown (through numerical simulation) to have low power for short series [11, 15].

Comparisons with other studies

Our findings align with previous simulation studies examining the performance of statistical methods for ITS. Huitema and McKean [40] similarly found that OLS confidence interval coverage decreased with increasing series length (with six lengths ranging from 12 to 500) in the presence of autocorrelation. McKnight et al. [14] similarly found that PW and ARIMA yielded liberal Type I error rates for the regression model parameters.

Other simulation studies have investigated the performance of methods for general time series, and our findings also align with these. Alpargu and Dutilleul [13] concluded from their simulation study examining the performance of REML, PW and OLS for lag (1) time series data over a range of series lengths (from 10 to 200), that REML is to be preferred over OLS and PW in estimating slope parameters. Cheang and Reinsel [20] examined the performance of ML and REML for estimating linear trends in lag (1) time series data of length 60 and 120 (both with and without seasonal components) and concluded that the REML estimator yielded better confidence interval coverage for the slope parameter, and less biased estimates of autocorrelation. Smith and McAleer [12] examined the performance of the NW estimator for time series of length 100 with lags of 1, 3 and 10, and found that it underestimated the SEs of the slope parameter.

Strengths and limitations

The strengths of our study include that we have used many combinations of parameter estimates and statistical methods. Our parameter values were informed by characteristics of real world ITS studies [4]. We planned and reported our study using the structured approach of Morris et al. [32] for simulation studies, and we generated a large number of data sets per combination to minimise MCSE.

As with all simulation studies, there are limitations to the applicability of findings. All data series were based on a random number generator and results may change given a different set of series, however, this is unlikely to be problematic given our MCSE was < 0.5% for all potential values of coverage and type I error rate. Our findings are only applicable to the scenarios in which they were generated, and so may not apply to ITS studies with different characteristics, such as unequal numbers of time points in the pre- and post-interruption segments, non-constant variance or different lags of autocorrelation (including seasonal effects).

Implications for practice

We found that all methods yielded unbiased estimates of the level and slope change, however, the methods differed in their performance in terms of confidence interval coverage and estimation of the autocorrelation parameter. Confidence interval coverage was primarily determined by the length of the time series and the underlying magnitude of autocorrelation. In practice, however, most analysts will only have knowledge of the length of the time series to guide in the choice of method. In rare cases, knowledge of the likely size of the underlying autocorrelation may be available from a previous long time series study in a similar context, which could help inform their choice. In our review of ITS studies investigating public health interruptions or exposures, the magnitude of autocorrelation was almost never explicitly specified (1%, 3/230 time series) [4]. Analysis of data extracted from the ITS studies included in this review using the REML method yielded a median autocorrelation 0.2 (IQR: 0 to 0.6, $n = 180$); however, as shown from the simulation study, the estimates of autocorrelation (on which these summary statistics are based) are likely to be underestimated.

From the statistical methods and scenarios we examined, we found that for small time series (approximately 12 points or under), in the absence of a method that performs well adjusting for autocorrelation in such short series, OLS is the recommended method. For longer time series, REML is recommended. If the analyst has knowledge that the underlying autocorrelation is likely to be large, then using REML with the Satterthwaite adjustment may be advantageous. However, when the Satterthwaite adjustment yields d.f. lower than 2, we recommend replacing these with 2 to mitigate nonsensible confidence intervals. When REML doesn't converge, ARIMA provides a reasonable alternative as, with the exception of REML, it yields higher confidence interval coverage than the other methods. Given most methods will yield confidence intervals that are too small, with type I error rates greater than 5%, borderline findings of statistical significance for the regression parameters should be cautiously interpreted; these may be due to chance rather than as a result of the interruption.

Estimates of autocorrelation from long series can be useful to inform sample size calculations and analytical decisions in future studies. We recommend reporting the REML estimates of the autocorrelation coefficient when possible. We only recommend using the DW test for detecting underlying autocorrelation in long time series (longer than 100 data points) and recommend against its use as part of a two-stage or stepwise approach to determine whether to use a statistical method that adjusts for autocorrelation.

In terms of study design, we recommend using 24 data points at the very minimum. With this number of points, confidence interval coverage close to the nominal 95% level can be achieved using REML with the Satterthwaite adjustment (when underlying autocorrelation is between 0 and 0.6). With fewer data points, poor confidence interval coverage is likely, irrespective of method.

Implications for future research

Although we investigated the statistical methods most commonly observed in reviews of ITS studies [1–4, 6], there is scope for further research examining other statistical methods, such as robust methods [41] or Bayesian approaches where the uncertainty in the estimate of autocorrelation could be incorporated. We investigated one small-sample adjustment (Satterthwaite) though others, such as Kenward-Roger [42], which adds a correction to the SE of regression parameter estimates, could also be examined. Further investigation of how the methods perform for scenarios other than those we investigated would be valuable. For example, when there are unequal numbers of points pre- and post-interruption, lags greater than 1, and where the autocorrelation and error variance differ between the pre and post interruption periods.

Conclusion

We undertook a simulation study to examine the performance of a set of statistical methods to analyse continuous ITS data under a range of scenarios that included different level and slope changes, varying lengths of series and magnitudes of lag-1 autocorrelation. We found that all methods yielded unbiased estimates of the level and slope change, however, the magnitude of autocorrelation was underestimated by all methods. This generally led to SEs that were too small and confidence interval coverage that was less than the nominal level. The DW test for the presence of autocorrelation performed poorly except for long series and large underlying autocorrelation. Care is needed when interpreting results from all methods, given the confidence intervals will generally be too narrow. Further research is required to determine and develop methods that perform well in the presence of autocorrelation, especially for short series.

Abbreviations

ARIMA: Autoregressive integrated moving average; ARMAX: Autoregressive moving average exogeneous; CI: Confidence interval; CO: Cochrane-orcutt; d.f.: Degrees of freedom; DW: Durbin-Watson; GLS: Generalised least squares; HAI: Healthcare-associated infection; IQR: Inter-quartile range; ITS: Interrupted time series; OLS: Ordinary least squares; MCSE: Monte-Carlo standard error; NHMRC: National health and medical research council; NW: Newey-west; PW: Prais-winsten; REML: Restricted maximum likelihood; Satt: Sattherthwaite; SD: Standard deviation; SE: Standard error

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-021-01364-0.

Additional file 1: Appendices This file contains the appendices referenced in the text. STurner_ITS_Simulation_Appendices.docx

Additional file 2: Supplementary File 1 This file contains the supplementary graphs described in the text.. STurner_ITS_Simulation_Supplementary_File_1.docx

---

Authors' contributions
JEM and ABF conceived the study and all authors contributed to its design. SLT designed and wrote the computer code and ran and analysed the simulations. SLT wrote the first draft of the manuscript, with contributions from JEM. SLT, JEM, AK, ABF, MT contributed to revisions of the manuscript and take public responsibility for its content.

Availability of data and materials
The data and computer code used to analyse the motivating example, the computer code used to create and analyse the simulated data sets, and the computer code used to plot the graphs in the manuscript (including Supplementary File 1) are available in the figshare repository, https://doi.org/10.26180/13284329 [33].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Public Health and Preventive Medicine, Monash University, 533 St Kilda Road, Melbourne, Victoria, Australia. [2]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Carling Ave, Ottawa, Ontario 1053, Canada. [3]School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Laurier Ave E, Ottawa, Ontario 75, Canada.

References

1. Ramsay CR, Matowe L, Grilli R, Grimshaw JM, Thomas RE. Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies. Int J Technol Assess Health Care. 2003;19(4):613–23. https://doi.org/10.1017/S0266462303000576.
2. Jandoc R, Burden AM, Mamdani M, Lévesque LE, Cadarette SM. Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations. J Clin Epidemiol. 2015;68(8):950–6. https://doi.org/10.1016/j.jclinepi.2014.12.018.
3. Ewusie J, Soobiah C, Blondal E, Beyene J, Thabane L, Hamid J. Methods, applications and challenges in the analysis of interrupted time series data: a scoping review. J Multidiscip Healthc. 2020;13:411–23. https://doi.org/10.2147/JMDH.S241085.
4. Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, Cheng AC, et al. Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: a review. J Clin Epidemiol. 2020;122:1–11. https://doi.org/10.1016/j.jclinepi.2020.02.006.
5. Wagner AK, Soumerai SB, Zhang F, Ross-Degnan D. Segmented regression analysis of interrupted time series studies in medication use research. J Clin Pharm Ther. 2002;27(4):299–309. https://doi.org/10.1046/j.1365-2710.2002.00430.x.
6. Hudson J, Fielding S, Ramsay CR. Methodology and reporting characteristics of studies using interrupted time series design in healthcare. BMC Med Res Methodol. 2019;19(1):137. https://doi.org/10.1186/s12874-019-0777-x.
7. Kontopantelis E, Doran T, Springate DA, Buchan I, Reeves D. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. BMJ Br Med J. 2015;350(jun09 5):h2750. https://doi.org/10.1136/bmj.h2750.
8. Lopez Bernal J, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol. 2016:dyw098.
9. Penfold RB, Zhang F. Use of interrupted time series analysis in evaluating health care quality improvements. Acad Pediatr. 2013;13(6):S38–44. https://doi.org/10.1016/j.acap.2013.08.002.
10. Gebski V, Ellingson K, Edwards J, Jernigan J, Kleinbaum D. Modelling interrupted time series to evaluate prevention and control of infection in healthcare. Epidemiol Infect. 2012;140(12):2131–41. https://doi.org/10.1017/S0950268812000179.
11. Huitema BE, McKean JW. Identifying autocorrelation generated by various error processes in interrupted time-series regression designs. Educ Psychol Meas. 2007;67(3):447–59. https://doi.org/10.1177/0013164406294774.
12. Smith J, McAleer M. Newey-West covariance matrix estimates for models with generated regressors. Appl Econ. 1994;26(6):635–40. https://doi.org/10.1080/00036849400000034.
13. Alpargu G, Dutilleul P. Efficiency and validity analyses of two-stage estimation procedures and derived testing procedures in quantitative linear models with AR(1) errors. Commun Stat Simul Comput. 2003;32(3):799–833. https://doi.org/10.1081/SAC-120017863.
14. McKnight SD, McKean JW, Huitema BE. A double bootstrap method to analyze linear models with autoregressive error terms. Psychol Methods. 2000;5(1):87–101. https://doi.org/10.1037/1082-989X.5.1.87.
15. Huitema BE, McKean JW. An improved portmanteau test for autocorrelated errors in interrupted time-series regression models. Behav Res Methods. 2007;39(3):343–9. https://doi.org/10.3758/BF03193002.
16. Brennan SE, McDonald, S., Cheng, A., Reid, J., Allen, K., McKenzie, J.E. Systematic review of novel disinfection methods to reduce infection rates in high risk hospitalised populations. Monash University, Melbourne, Australia.: Prepared by Cochrane Australia for the National Health and Medical Research Council. ; 2017.
17. Hacek DM, Ogle AM, Fisher A, Robicsek A, Peterson LR. Significant impact of terminal room cleaning with bleach on reducing nosocomial Clostridium difficile. Am J Infect Control. 2010;38(5):350–3. https://doi.org/10.1016/j.ajic.2009.11.003.
18. Huitema BE. Analysis of covariance and alternatives statistical methods for experiments, quasi-experiments, and single-case studies. 2nd ed. Hoboken, N.J: Wiley; 2011.
19. Huitema BE, Mckean JW. Design specification issues in time-series intervention models. Educ Psychol Meas. 2000;60(1):38–58. https://doi.org/10.1177/00131640021970358.
20. Cheang W-K, Reinsel GC. Bias reduction of autoregressive estimates in time series regression model through restricted maximum likelihood. J Am Stat Assoc. 2000;95(452):1173–84. https://doi.org/10.1080/01621459.2000.10474318.
21. Kutner M, Nachtscheim C, Neter J, Li W, Senter H. Applied linear statistical models. In: Kutner M, Nachtscheim C, Neter J, Li W, Senter H, editors. 2008. p. 880-.
22. Newey WK, West KD. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica. 1987;55(3):703. https://doi.org/10.2307/1913610.
23. StataCorp. Stata 15 Base Reference Manual. College Station, TX: Stata Press; 2017.
24. Singer J, Willett J. Applied longitudinal data analysis: modeling change and event occurrence 2003. xx-xx p.
25. Thompson WA. The problem of negative estimates of variance components. Ann Math Stat. 1962;33(1):273–89. https://doi.org/10.1214/aoms/1177704731.
26. Satterthwaite FE. An approximate distribution of estimates of variance components. Biom Bull. 1946;2(6):110–4. https://doi.org/10.2307/3002019.
27. Paolella MS. Linear models and time-series analysis : regression, ANOVA, ARMA and GARCH. Hoboken, NJ: John Wiley & Sons, Inc; 2019.
28. Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, McKenzie JE. Comparison of six statistical methods for interrupted time series studies: empirical evaluation of 190 published series. BMC Med Res Methodol. 2021; 21(1):134. https://doi.org/10.1186/s12874-021-01306-w.
29. Nelson BK. Statistical methodology: V. Time series analysis using autoregressive integrated moving average (ARIMA) models. Acad Emerg Med. 1998;5(7):739.
30. Box GEPa. Time series analysis : forecasting and control. 5th ed: Hoboken, New Jersey : Wiley; 2016.
31. Durbin J, Watson GS. Testing for serial correlation in least squares regression: I. Biometrika. 1950;37(3/4):409–28.
32. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074–102. https://doi.org/10.1002/sim.8086.
33. Turner SL, Forbes AB, Karahalios A, Taljaard M, McKenzie JE. Evaluation of statistical methods used in the analysis of interrupted time series studies: a simulation study - code and data 2020.
34. White IR. Simsum: analyses of simulation studies including Monte Carlo error. Stata J. 2010;10:369.
35. Stata. Stata Statistical Software, vol. 15. College Station, TX: Statcorp LLC; 2017.
36. Box GEP, Pierce DA. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J Am Stat Assoc. 1970;65(332):1509–26. https://doi.org/10.1080/01621459.1970.10481180.
37. Ljung GM, Box GEP. On a measure of lack of fit in time series models. Biometrika. 1978;65(2):297–303. https://doi.org/10.1093/biomet/65.2.297.
38. Breusch TS. Testing for autocorrelation in dynamic linear models. Aust Econ Pap. 1978;17(31):334–55. https://doi.org/10.1111/j.1467-8454.1978.tb00635.x.
39. Godfrey LG. Testing against general autoregressive and moving average error models when the Regressors include lagged dependent variables. Econometrica. 1978;46(6):1293–301. https://doi.org/10.2307/1913829.
40. Huitema BE, McKean JW, McKnight S. Autocorrelation effects on least-squares intervention analysis of short time series. Educ Psychol Meas. 1999; 59(5):767–86. https://doi.org/10.1177/00131649921970134.
41. Cruz M, Bender M, Ombao H. A robust interrupted time series model for analyzing complex health care intervention data. Stat Med. 2017;36(29): 4660–76. https://doi.org/10.1002/sim.7443.
42. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics. 1997;53(3):983–97. https://doi.org/10.2307/2533558.

Publisher's Note