Change detection under autocorrelation

Maarten Speekenbrink (m.speekenbrink@ucl.ac.uk), Matthew A. Twyman (m.twyman@ucl.ac.uk)
Nigel Harvey (n.harvey@ucl.ac.uk)
Cognitive, Perceptual and Brain Sciences, University College London
Gower Street, London WC1E 6BT, England

Abstract

Judgmental detection of changes in time series is an ubiquitous task. Previous research has shown that human observers are often relatively poor at detecting change, especially when the series are serially dependent (autocorrelated). We present two experiments in which participants were asked to judge the occurrence of changes in time series with varying levels of autocorrelation. Results show that autocorrelation increases the difficulty of discriminating change from no change, and that observers respond to this increased difficulty by biasing their decisions towards change. This results in increased false alarm rates, while leaving hit rates relatively intact. We present a rational (Bayesian) model of change detection and compare it to two heuristic models that ignore autocorrelation in the series. Participants appeared to rely on a simple heuristic, where they first visually match a change function to a series, and then determine whether the putative change exceeds the variability in the data.

Keywords: change detection; judgment; forecasting

Introduction

Detecting changes in time series is a surprisingly ubiquitous task. Doctors and therapists monitor diagnostic indicators for signs of disease onset and for evidence that a prescribed treatment is effective; farmers monitor soil conditions to decide whether additional irrigation is necessary; local authorities monitor river levels for increased likelihood of flooding; probation officers monitor probationers' behaviour for evidence of return to crime; financiers monitor data, such as exchange rates, for signs of trend reversal. Many other examples could be given. As with forecasting and control tasks, monitoring tasks may be tackled by formal statistical methods, by using judgment alone, or by using some combination of these two approaches. The method most favoured depends to a large extent on the domain. Typically, implementation of and training in formal methods consume more resources (time, money, effort) but the investment may be worthwhile if those methods have considerable benefits over judgment in terms of accuracy. Thus, it would be useful to know just how good human judgment is relative to formal methods.

There are many formal statistical methods for detecting change in time series (e.g., Albert & Chib, 1993; Carlin, Gelfand, & Smith, 1992; Hamilton, 1990). This variety is partly because some approaches represent the event producing the regime change as deterministic whereas others represent it as a random variable and partly because, whichever of these approaches is adopted, there is still some debate about how best to estimate the likelihood that a change has occurred.

In contrast, there has been very little research into judgmental assessment of regime change. Originally, behavioural psychologists working within the Skinnerian tradition used judgment (visual inference) to assess whether a manipulation changed some aspect of an animal's behaviour represented as a time series. They argued that this is a conservative approach because only large effects can be detected (e.g., Baer, 1977). Their claims were not directly tested. However, when behaviour analysts later used the same approach to assess human patients, there was concern that the shorter pre-treatment baselines in the series impaired visual inference. As a result, some experiments were carried out to investigate how accurately people can detect change.

Judgmental change detection and autocorrelation

Jones, Weinrott, and Vaught (1978) found that people were poor at detecting change in real series: inter-rater reliability of judgments was low at .39 and average miss and false alarm rates were 48% and 33%, respectively. Sequential dependence (autocorrelation) in series increased false alarm rates. This study used interrupted time series analysis as the gold standard for establishing whether there was a real change in the series. However, series were so short that this statistical approach would have lacked power. People may have been able to detect changes that the statistical analysis could not: if so, their performance may not have been as bad as it appeared to be. To circumvent this problem, Matyas and Greenwood (1990) simulated series with known levels of random noise and first-order autocorrelation. However, they still found that false alarm rates (typically over 40%) were much higher than miss rates (typically about 10%), especially when data were autocorrelated. They concluded that judgment is not as conservative as behaviour analysts assumed.

The increase in false alarm rates under positive autocorrelation is problematic. In single-subject research, where visual assessment of change is still the dominant method (Brossart, Parker, Olson, & Mahadevan, 2006), there is positive autocorrelation in the large majority of series (Busk & Marascuilo, 1988). Why does autocorrelation impair change detection? Consider a time series $y_{1:T} = (y_1, \ldots, y_T)$ which follows an r-th order autoregressive process

$$y_t = \mu_t + \sum_{k=1}^{r} \alpha_k (Y_{t-k} - \mu_{t-k}) + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \tag{1}$$

This process implies a serial dependence between successive time points such that when a previous value y_{t-k} is above the mean μ_{t-k}, a later value y_t is more likely to also be above the mean (for $\alpha_k > 0$, positive autocorrelation), or more likely to be below the mean ($\alpha_k < 0$, i.e., negative autocorrelation).

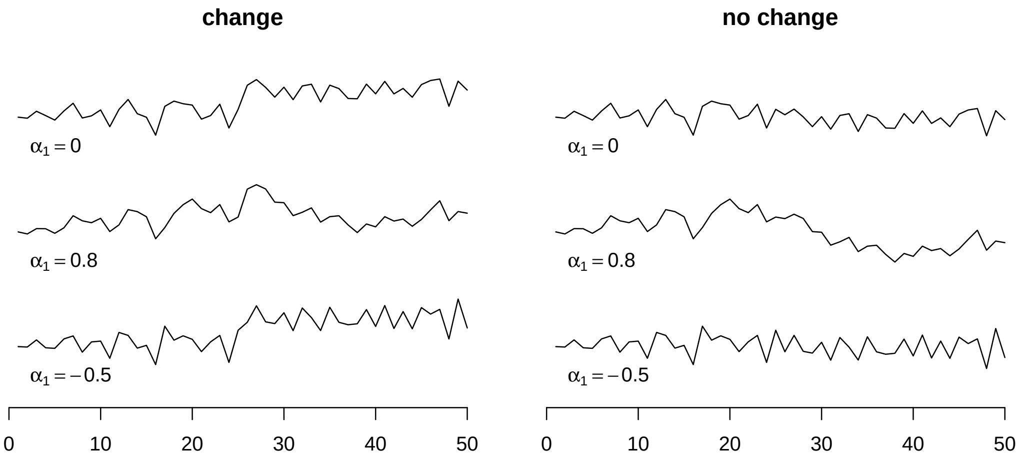

Figure 1: Examples of time series with a change and no change, under different levels of autocorrelation.

We allow for the possibility of an abrupt change in baseline value

$$\mu_t = \mu_0 + \delta \mathbb{I}_{t \geq t^*} \qquad \delta \sim N(\mu_\delta, \sigma_\delta^2) \tag{2}$$

where t^* is the change point and $\mathbb{I}_{t \geq t^*}$ is an indicator function with value 1 when $t \geq t^*$ and value 0 otherwise. This is the process used in the experiments reported below. Figure 1 shows examples of time series produced by a first-order ($r = 1$) autoregressive process. Note that each series is created from identical noise values ε_t; the only difference is in the value of the autocorrelation, and whether there is a change in the baseline (after time point 25 in the series on the left) or not. As can be seen in these examples, positive autocorrelation ($\alpha_1 = 0.8$) tends to make series "smoother", which can make abrupt changes less apparent. Also, when there is no change, positive autocorrelation can increase false alarms, because a sudden large noise perturbation will tend to persist, giving the appearance of a change in mean. In contrast, negative autocorrelation ($\alpha_1 = -.5$) tends to make series more "jagged", which can impair change detection by increasing the apparent noisiness of the series, even though the actual error variance is unaffected.

Cognitive processes underlying change detection

Research on judgmental change detection has been mainly empirical. Little attention has been given to the cognitive processes underlying performance in the task. An exception is the work of Wampold and Furlong (1981), who argued that people may use one of two types of heuristic.

First, people may assess a putative change in the data relative to the overall variability that is present. A putative change is judged to be real if its magnitude exceeds the natural variability in the data by some criterial amount. This is a global assessment: all the presented data are taken into account. According to this model, positive autocorrelation increases false alarms because it is ignored when the natural variability in the data is estimated. This results in an underestimation of the variability and, hence, over-estimation of the likelihood of a change. In contrast, negative autocorrelation would result in an over-estimation of the variability, and hence an underestimation of the likelihood of a change.

Alternatively, people may monitor the data for large absolute changes and ignore the natural point-to-point variability. This approach assumes that they access internal prototypes of possible changes and classify series into those with and without a change by matching them against these prototypes. In other words, they make local assessments: candidate changes are judged in isolation from the rest of the series.

Steyvers and Brown (2005) proposed that people may rely on a heuristic similar to the latter[1] when detecting changes in times series while making sequential predictions about the next datum in the series. They showed that this heuristic closely followed an optimal Bayesian model, which also fitted participants responses well. In later work, Brown and Steyvers (2009) proposed that people generally use Bayesian inference in these problems, although they may work from incorrect assumptions and use limited samples to approximate the full Bayesian analysis.

Present study

Previous research suggests that human detection of change in time series is impaired by the presence of autocorrelation. In the two experiments presented here, we sought to replicate this finding with a range of (positive) autocorrelations (Experiment 1) and a second-order autoregressive process (Experiment 2). By formulating different models and fitting these to participants' judgments, we sought to uncover the cognitive processes underlying change detection in graphically

[1] But not exactly the same. According to Steyvers and Brown (2005), a change will be detected whenever the distance between a prediction and the actual value exceeds a criterial amount. However, the distance was not measured on the scale of the outcome, but rather as the log likelihood ratio of the outcome given a change vs no change.

presented time series. Uncovering the strategies people use when monitoring series for regime change is a first step in determining how to improve judgmental change detection.

Experiment 1

The objective of Experiment 1 was to assess the accuracy of judgmental change detection under levels of no ($\alpha_1 = 0$), medium ($\alpha_1 = .4$), and high ($\alpha_1 = .8$) autocorrelation.

Method

Participants Fifty participants (23 male) were recruited from the UCL subject pool and paid £5 for their time. The mean age was 26.06.

Task The change detection task consisted of 60 trials. On each trial, participants were presented a graph depicting a time series and asked to indicate whether the series contained a change or not. After this, they indicated their confidence in their response.

In Experiment 1, the time series were created by a first-order autoregressive process (i.e., setting $r = 1$ in Equation 1). The autocorrelation was varied within participants. For each level of autocorrelation, $\alpha_1 = \{0, .4, .8\}$, there were 10 series with, and 10 series without a change. Each change δ was randomly drawn from a Gaussian distribution with mean $\mu_\delta = 8$ and variance $\sigma^2_\delta = 9$. The initial baseline value was set at $\mu_0 = 50$ and the variance of the noise was set at $\sigma^2_\varepsilon = 5$.

Procedure Participants took the role of a trainee flood engineer with the task of monitoring water levels for risk of flooding. Participants were told a risk of flooding consisted of a persistent increase in water level, but that the level would fluctuate regardless of whether there was a risk of flooding or not. For 60 different locations, they would monitor the water level over a 50 hour period, and participants were informed that a flood risk could occur anywhere between hour 11 and hour 40 (i.e., $t^* \in \{11, \ldots, 40\}$). Finally, participants were instructed that there was a flood risk for half of the locations.

Results

The main detection results are given in Table 1. Autocorrelation did not affect hit rates, $F(2,98) = 1.278$, $p = .283$, but increased false alarms, $F(2,98) = 27.913$, $p < .001$. This indicates that, as autocorrelation increased, participants adjusted their criterion to detect changes. This was confirmed in a signal detection analysis. Increased levels of autocorrelation reduced the discrimination (d'), $F(2,98) = 11.915$, $p < .001$. Contrast analysis showed that this effect was mainly due to a linear trend, $F(1,49) = 22.79$, $p < .001$; the quadratic trend was not significant, $F(1,49) = 1.37$, $p = .25$. In addition, increased autocorrelation reduced the centered decision criteria (C), $F(2,98) = 20.56$, $p < .001$. Contrast analysis showed that this effect was mainly due to a linear trend, $F(1,49) = 32.58$, $p < .001$; the quadratic trend was not significant, $F(1,49) = 0.79$, $p = .38$. This indicates that autocorrelation increased the difficulty of the task and participants responded by lowering the criterion to detect a change (biasing decisions towards changes), resulting in increased false alarms.

Table 1: Mean hit (H) and false alarm (FA) rates, discrimination (d') and (centered) criterion (C) parameters. Values in parentheses are standard deviations.

α_1	α_2	H	FA	d'	C
		Experiment 1			
0		.72 (.18)	.06 (.15)	2.06 (.69)	-0.40 (.12)
0.4		.76 (.14)	.11 (.19)	1.90 (.65)	-0.44 (.12)
0.8		.77 (.15)	.24 (.20)	1.48 (.77)	-0.50 (.11)
		Experiment 2			
0.5	0.3	.68 (.24)	.14 (.15)	1.65 (.86)	-0.42 (.14)
0.5	0.0	.74 (.22)	.10 (.15)	2.01 (.85)	-0.43 (.12)
0.5	-0.3	.70 (.23)	.07 (.16)	2.00 (.81)	-0.40 (.14)
-0.5	0.3	.72 (.22)	.07 (.14)	2.07 (.89)	-0.41 (.12)
-0.5	0.0	.72 (.23)	.03 (.07)	2.18 (.67)	-0.39 (.12)
-0.5	-0.3	.68 (.24)	.03 (.09)	2.09 (.80)	-0.37 (.12)

Average confidence levels for the different trial types and autocorrelation levels are depicted in Figure 2. We analysed the confidence ratings with a linear mixed effects model, including random intercepts for each participant, as well as random slopes for the autocorrelation and contrast codes for whether a decision was correct and whether it was a "change" (vs a "no change") decision. This showed that confidence was higher for correct (hits and correct rejections) than incorrect decisions (misses and false alarms), $F(1,2946) = 216.4$, $p < .001$. In addition, confidence was generally lower when participants responded change (hits and false alarms) compared to no change (correct rejections and misses), $F(1,2946) = 11.19$, $p < .001$. A significant interaction between these two factors shows that confidence was more strongly related to correctness when people judged there was a change than when people judged there was no change, $F(1,2946) = 4.68$, $p = .031$. Finally, confidence decreased as the level of autocorrelation increased, $F(1,2946) = 16.83$, $p < .001$.

To summarize the results, increasing autocorrelation resulted in poorer discrimination between series with and those without a change. This increased difficulty was also reflected in participants' confidence in their judgments. Participants appeared to respond to the increased difficulty by relaxing their decision criteria in favour of detecting change, resulting in increased false alarm rates.

Experiment 2

The objective of Experiment 2 was to investigate change detection in a second-order autoregressive process. Depending on the autocorrelation values, second-order autoregressive processes can show complex periodic patterns (e.g., Gottman, 1981). In particular, when $\alpha_1^2 + 4\alpha_2 < 0$, the spectral density functions show broad peaks across a band of mid-range

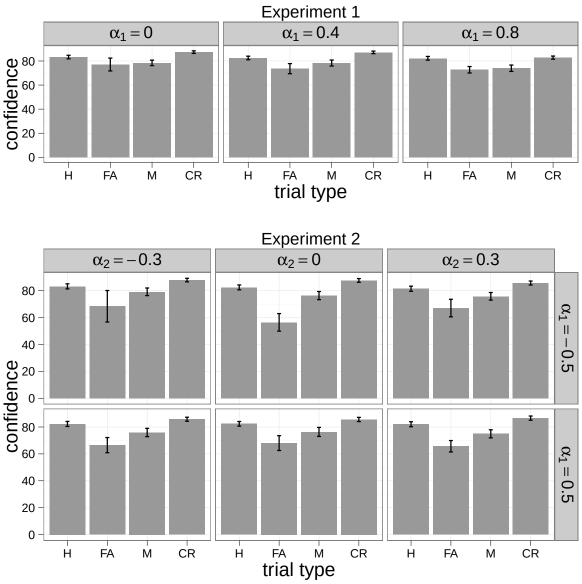

Figure 2: Mean confidence levels by autoregression level and trial type (H = hit, FA = false alarm, M = miss, CR = correct rejection). Error bars represent 95% confidence intervals.

frequencies, and the time series is nondeterministically periodic. Such periodic behaviour may appear as a change in mean, increasing false alarms. To investigate this possibility, Experiment 2 included series with and without such periodic behaviour.

Method

Participants 70 students (18 male) participated in the experiment for course credit. The mean age was 18.97 (*SD* = 1.12).

Task The task was identical to that in Experiment 1. However, the time-series were now produced by a second-order autoregressive process (setting $r = 2$ in Equation 1), with autocorrelation parameters $\alpha_1 = \{-0.5, 0.5\}$ and $\alpha_2 = \{-0.3, 0, 0.3\}$. There were 10 series for each combination of the autocorrelation parameters, of which five did and five did not contain a change. Series were presented in random order. Series with $\alpha_1 = .5$ and $\alpha_2 = -.3$, or $\alpha_1 = -.5$ and $\alpha_2 = -.3$, were likely to show periodic trends. If this impairs detection ability, we would expect to find a difference between these series and those generated with other combinations of autocorrelation parameters.

Results

The main detection results can be found in Table 1. As in Experiment 1, autocorrelation affected false alarm rates, $F(1,68) = 56.67$, $p < .001$ and $F(2,136) = 24.99$, $p < .001$, for α_1 and α_2 respectively, but not hit rates, $F(1,68) = 0.03$, $p = .87$ and $F(2,136) = 1.42$, $p = .25$. Signal detection analysis showed that both autocorrelation parameters affected discrimination ability (d'), $F(1,69) = 9.81$, $p = .003$, and $F(2,138) = 3.50$, $p = .033$, for α_1 and α_2 respectively; the interaction was not significant, $F(2,138) = 1.79$, $p = .17$. Discrimination was generally better for $\alpha_1 = -.5$ than $\alpha_1 = 0.5$. Discrimination was worse for $\alpha_2 = 0.3$ compared to the other two values, $F(1,69) = 5.77$, $p = .019$, while there was no difference between $\alpha_2 = 0$ and $\alpha_2 = -0.3$, $F(1,69) = 0.52$, $p = .47$. Both autocorrelation parameters also affected the centered decision criteria (C), $F(1,69) = 7.52$, $p = .008$, and $F(2,138) = 3.17$, $p = .045$. Decision criteria were more biased towards change for positive compared to negative autocorrelations (for α_2, there was a linear trend, $F(1,69) = 5.78$, $p = .019$, but no quadratic trend, $F(1,69) = 0.81$, $p = .37$).

Average confidence levels for the different trial types and autocorrelation levels are depicted in Figure 2. Analysis with a linear mixed effects model replicated the results of Experiment 1. Confidence was higher for correct (hits and correct rejections) than incorrect decisions (misses and false alarms), $F(1,4124) = 226.14$, $p < .001$, and lower when participants responded change (hits and false alarms) compared to no change (correct rejections and misses), $F(1,4124) = 44.80$, $p < .001$. Confidence was more strongly related to correctness when people judged there was a change than when people judged there was no change, $F(1,4124) = 8.1$, $p = .005$. Confidence was higher for negative than positive first-order autocorrelation (α_1), $F(1,4124) = 5.69$, $p = .017$, and this difference was larger when the second-order autocorrelation was negative rather than positive, $F(1,4124) = 5.62$, $p = .018$.

The results of this experiment replicate those of Experiment 1 with a second-order autoregressive process. Positive (first- and second-order) autocorrelation resulted in poorer discrimination between series with and series without a change and this increased difficulty was reflected in participants' confidence in their judgments. Participants responded to the increased difficulty by relaxing their decision criteria, increasing false alarm rates. As the interaction between the autocorrelations was not significant, we found no evidence that periodic trends in the time series affected detection ability.

A Bayesian change detection model

To compare participants' judgments against a gold standard, we used a Bayesian model to detect changes in time series as defined by Equations 1 and 2. By taking the autocorrelation into account, this model is expected to perform well, although the relatively short length of the series may limit its performance.

In our analysis, change detection is based on the relative evidence for a model M_1, which incorporates the possibility of a change, over a model M_2, which does not allow for change. The measure of relative evidence is the Bayes Factor

$$BF = \frac{p(y_{1:T}|M_1)}{p(y_{1:T}|M_2)} \tag{3}$$

where $p(y_{1:T}|M_j)$ the marginal likelihood

$$p(y_{1:T}|M_j) = \int p(y_{1:T}, \theta_j|M_j)d\theta_j \tag{4}$$

integrating over the parameters θ_j of model j. For M_1, the parameters are $\theta_1 = \{\mu_0, \delta, \alpha_1, \alpha_2, \sigma_\varepsilon, t^*\}$. The parameters of model M_2 exclude δ and t^*. Recall that participants were informed there could be only one change point in each series, and that possible change points could be anywhere between time points 11 and 40. In model 1, the posterior distribution of the change point $t^* \in \{11,\ldots,40\}$, conditional on the other parameters, can then be expressed relatively simply as

$$p(t^*|y_{1:T}, \theta_{1,-t^*}, M_1) \propto p(y_{1:T}|t^*, \theta_{1,-t^*}, M_1)p(t^*|M_1) \tag{5}$$

where $\theta_{1,-t^*}$ denotes the parameter vector excluding t^*, and $p(t^*|M_1)$ the prior distribution over the change points t^*, which we took to be uniform. For μ_0 and δ, we used truncated Normal distributions with means of 50 and 0 respectively, and variance 10^5. Both distributions were restricted to the range between 0 and 100. For μ_0, this reflects the range of the time series on the graphs. For δ, this reflects that changes can only be positive. For σ_ε, an inverse Gamma distribution was used. For α_1 and α_2, we used truncated Normal distributions (centered on 0) restricting the range such that the process is stationary[2]. Posterior distributions for the parameters can be efficiently estimated by Gibbs sampling (for computational details, see Albert & Chib, 1993). Bayes Factors were estimated from the Gibbs sampler using the technique of Chib (1995).

We computed the Bayes Factor for each of the time series in the two experiments. Using the simple criterion of $BF > 1$ to detect a change, the hit rate of the model was 92.3%, but the false alarm rate was rather high at 15.8%. Inspection of the parameter estimates showed that false alarms were generally associated with relatively small changes (50% of the posterior means of δ were smaller than 2.31) occurring relatively late in the series (50% of the posterior modes of $p(t^*)$ were larger than $t = 33$). False alarm rates increased with positive autocorrelation. This suggests that even while explicitly accounting for autocorrelation, the Bayesian model is not immune to illusory changes produced by autocorrelation.

Table 2: Model fits (AIC) and numbers of participants best fitted according to the AIC (n_{best}).

	Experiment 1		Experiment 2	
	AIC	n_{best}	AIC	n_{best}
Bayes	2891	3	3728	9
CRV	2280	47	2846	61
LAC	3512	0	5314	0

Modelling human change detection

To link the Bayesian and heuristic models to participants' responses R_{ik}, we assume that, for a time-series k, each method of change detection j provides a signal v_{jk}, which is corrupted by noise $e_{ijk} \sim N(0, \sigma_{ij})$, and that participants judge there to be a change when the noisy signal exceeds a criterion c_{ij}. As a result, the probability of a change judgment, for participant i judging series k with model j, can be written as

$$P(R_{ik} = \text{change}|j) = 1 - \Phi\left(\frac{v_{jk} - c_{ij}}{\sigma_{ij}}\right) \tag{6}$$

where Φ denotes the cumulative Normal distribution. For the Bayesian model, we assume that

$$v_{\text{Bayes},k} = \log BF \tag{7}$$

i.e., that decisions depend on the logarithm of the Bayes Factor (Equation 3).

In addition to the Bayesian model, we formalised the heuristics of Wampold and Furlong (1981) described in the introduction. We'll refer to these as the Change-Relative-to-Variability (CRV) and the Largest-Absolute-Change (LAC) heuristic. The CRV heuristic compares a putative change to the overall variation in the series. In doing so, we assume people first visually fit a step function to the series, in such a way as to minimize the noise. The step function represents the mean water level before and after the change-point, and with that the timing and size of the putative change in water level. This putative change is compared to the deviations of the series around the step function. More formally, we assume the (increasing) step function, defined by the initial level m_0, change point t', and the increase d after the change, is determined by minimizing the sum of squared error (SSE):

$$t', m_0, d = \underset{t', m_0, d}{\arg\min} \sum_{t=1}^{T} (Y_t - m_0 - d\mathbb{I}_{t \geq t'})^2 \tag{8}$$

The putative change d is then compared to an estimate s of the standard deviation (derived from the sum of squared errors). The value used for decisions, according to this heuristic, is then simply

$$v_{CRV,k} = d - s \tag{9}$$

According to the second heuristic, a change is determined solely by comparing deviations between time points to a preexisting "prototype". For this heuristic, we therefore assume the signal consists of the maximum deviation between consecutive time points

$$v_{LAC,k} = \max_t |y_t - y_{t-1}| \tag{10}$$

We fitted the models in two ways. First, we fitted each model to the whole group of participants using a generalized linear mixed effects model, including random decision criteria c_{ij} and dispersion parameters σ_{ij}. In addition, we fitted each model to each participant separately. The results of both analyses (Table 2) were in agreement: the model that best described participants' responses was the Change-Relative-to-Variability (CRV) heuristic, followed by the Bayesian model.

[2] For a first-order autoregressive process, that means that $|\alpha_1| < 1$. For a second-order autoregressive process, the requirements are that $\alpha_1 + \alpha_2 < 1$, $\alpha_2 - \alpha_1 < 1$, $|\alpha_2| < 1$.

The Largest-Absolute-Change (LAC) heuristic fitted none of the participants best.

Discussion

We presented two experiments on human change detection in autocorrelated time series. In agreement with Matyas and Greenwood (1990), we found little evidence of conservatism. As the level of autocorrelation increased, participants maintained a similar hit rate, but increased their false alarm rate, indicating a relaxing of decision criteria such that more changes are (erroneously) detected. For second-order autoregressive processes, detection was most impaired when both autocorrelations were positive. There was no evidence that periodic trends resulting from particular autocorrelation levels affected detection performance.

Most participants responded in accordance with a simple change detection heuristic, where an underlying change-in-mean function is (visually) fitted to a noisy series and it is determined whether the putative change exceeds the natural variability in the series.[3] For the graphically presented time series used here, this strategy seems plausible. And as this strategy closely matches the performance of the Bayesian analysis which accounts for autocorrelation in the series, it is not a bad strategy either – at least not for the types of series studied here. Participants were explicitly told the range of time points over which a change could occur, as well as that there could only be one change in each series. For more complex problems with multiple change points or changes in other parameters than the baseline, the Bayesian and heuristic analyses are more likely to diverge.

We found no evidence that people relied on (absolute) differences between values on successive time points, a heuristic suggested by e.g. Steyvers and Brown (2005). An important difference between the present study and the latter one is that our participants were presented with complete time series, while Steyvers and Brown used an online prediction task in which participants viewed each datum sequentially and thus had to rely on memory. It is likely that change detection strategies will differ between online and offline detection tasks. In online tasks, the information available for detection is constrained by (working) memory capacity, as well as by the fact that only previous data can be used to judge a change at the current datum. In this case, strategies that rely on small samples, such as an absolute change heuristic, seem more plausible than strategies which use all the data, such as the Bayesian model and Change-Relative-to-Variability heuristic as implemented here.

Autocorrelation clearly impeded detection performance. Further research is required to assess the extent to which people can learn to "see through" autocorrelation. In the present experiments, participants did not receive feedback about their detection performance. It is possible that, after extensive training, people can learn to distinguish between real changes and those that are merely apparent due to autocorrelation. In domains such as risk assessment, the finding that autocorrelation increases false alarms, but does not decrease hit rates, may provide some comfort. However, when assessing treatment effectiveness, a more cautious approach may be called for.

Acknowledgements

This work was supported by the Economic and Social Research Council (ESRC), grant RES-062-23-2735.

References

Albert, J. H., & Chib, S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business and Economic Statistics*, *11*, 1–15.

Baer, D. M. (1977). Perhaps it would be better not to know everything. *Journal of Applied Behavior Analysis*, *10*, 167–172.

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification*, *30*, 531–563.

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive psychology*, *58*, 49–67.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment*, *10*, 229–242.

Carlin, B. P., Gelfand, A. E., & Smith, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Journal of the Royal Statistical Society. Series C (Applied statistics)*, *41*, 389–405.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*, 1313–1321.

Gottman, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientist.* Cambridge: Cambridge University Press.

Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics*, *45*, 39–70.

Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis*, *10*, 151–166.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single case time-series: Effects of variability, serial dependence and magnitude of intervention effects. *Journal of Applied Behavior Analysis*, *23*, 341–351.

Steyvers, M., & Brown, S. D. (2005). Prediction and change detection. In *Advances in neural information processing systems, 18* (pp. 1281–1288).

Wampold, B., & Furlong, M. (1981). The heuristics of visual inference. *Behavioral Assessment*, *3*, 79–82.

[3] In the present experiment, the variability actually seemed to have little effect on responses and a version without variability fitted just as well. However, as differences in variability between the series were relatively small, further research is required.