A changepoint approach to modelling soil moisture dynamics

Mengyi Gong, Rebecca Killick, Christopher Nemeth

Department of Mathematics and Statistics, Lancaster University, Lancaster, U.K.

COMPSTAT 2022, Bologna, Italy

Soil moisture time series from underground sensors

Underground sensors measure the volumetric water content of the soil at a high frequency. The large number of observations bring challenges to the modelling of soil infiltration features.

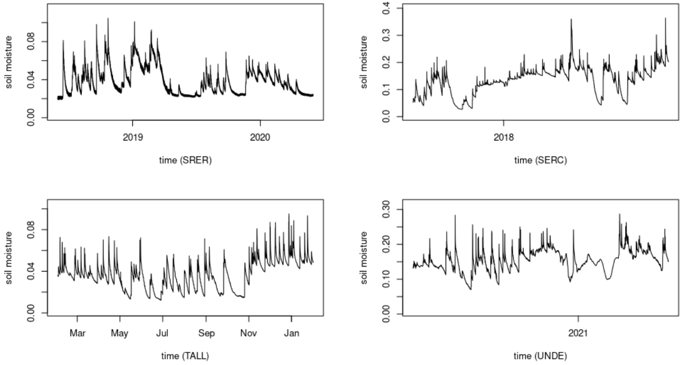

Figure: 30-minute soil moisture time series from NEON data portal.

Soil moisture drydown modelling

Soil moisture drydown modelling is commonly used in soil science to learn the infiltration features.
The typical modelling process requires manually separating the soil moisture time series into segments and fitting exponential decay models to them.
This can be time consuming.

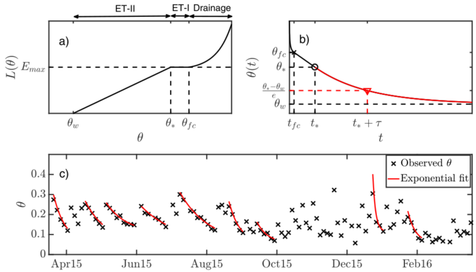

Figure: The soil mositure (θ) loss $L(\theta)$, the soil drydown curve $\theta(t)$ and the SMAP soil moisture observations of a year. The figure is taken from [McColl *et al.*(2017)].

The motivation of the changepoint-based approach

Motivated by the problem of recovering the underlying spike train from the noisy calcium fluorescence trace data in neoroscience [Jewell & Witten(2018), Jewell *et. al.*(2020)], we proposes a changepoint-based approach to automatically identify structural changes in the soil drying process.

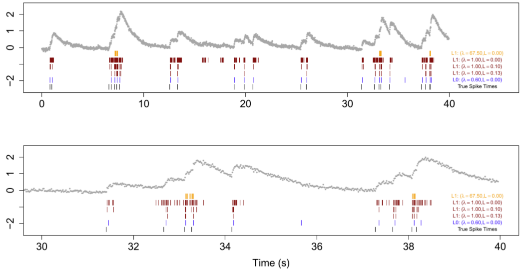

Figure: The flourence trace data from cell 2002 (grey), the true spike times (black) and the spikes detected by different methods (red, orange, blue). The figure is taken from [Jewell & Witten(2018)].

The motivation of the changepoint-based approach

Following an event (e.g. strom), the soil water content would rise suddenly, then decrease exponentially towards an asymptotic level, until the next event disrupts the decay process, leading to the next rise in soil moisture.

- The **time point where the sudden rise occurs**, or a time point within a short window of the sudden rise, **is considered as a changepoint**. The time series following the changepoint is assumed to follow an **exponential decay process**.
- The parameters in the exponential decay model are assumed to **vary from segment to segment**. This helps to investigate the temporal dynamics of soil moisture.

The 1st changepoint-based approach

- The goal is to **identify the changepoints** $\tau_i \in \{1, \cdots, n-1\}$, $i = 1, \cdots, k$, where the sudden rises occur and the exponential decays are disruptted.
- This is to minimise the **penalised cost function**

$$\sum_{i=0}^{k} \mathcal{C}(Y_{(\tau_i+1):\tau_{i+1}}) + \lambda f(k) \tag{1}$$

with respect to τ_i, $i = 1, \cdots, k$.
- The cost function is (two times) the negative log-likelihood of the **exponential decay model** fitted to the segment between τ_i and τ_{i+1},

$$Y_t = \alpha_{fi} + \alpha_{0i} \exp[-\exp(\gamma_i)\,(t - \tau_i)] + \epsilon_t\,, \tag{2}$$

where α_{fi} is the asymptotic soil moisture level, α_{0i} is size of the rise, and γ_i is the exponential decay parameter for segment i.

Identify the changepoints with PELT

- We developed **an algorithm based on PELT** to minimise the cost function (1).
- Starts with the recursive computation of the cost functions

$$F(s) = \min_{0 \leq \tau < s} \{F(\tau) + \mathcal{C}(Y_{(\tau+1):s}) + \lambda\}$$

- The PELT pruning criterion [Killick *et. al.*(2012)]: for all $t < t' < s$ satisfying

$$\mathcal{C}(Y_{(t+1):t'}) + \mathcal{C}(Y_{(t'+1):s}) + K \leq \mathcal{C}(Y_{(t+1):s}) \tag{3}$$

for some constant K, the time point t can never be the last optimal changepoint prior to time point s if $F(t) + \mathcal{C}(Y_{(t+1):t'}) + K \geq F(t')$.

Identify the changepoints with PELT

- For a model with multiple parameters, methods such as the functional pruning can encounter problem when identifying the multi-dimensional region where the cost function attains its minimum, undermining the computational efficiency.
- Using (two times) the negative log-likelihood of exponential decay model as the cost function satisfies the inequality (3) with $K = 0$.
- The proposed method can be implemented in R, with the non-linear least squares estimation carried out using existing R packages, e.g. nlmrt.

Simulation study

A simulation study was carried out to access the performance of the proposed method. We simulated **three different scenarios for the changepoints**, which are

1. sudden rises randomly distributed over time
2. sudden rises following a temporal pattern where one part of the time series has more frequent rises
3. large scale sudden rises randomly distributed over time, along with small scale rises over a long decay process

Each of the scenarios was then paired with **two noise levels**. We then run each scenario for 200 times.

Simulation study

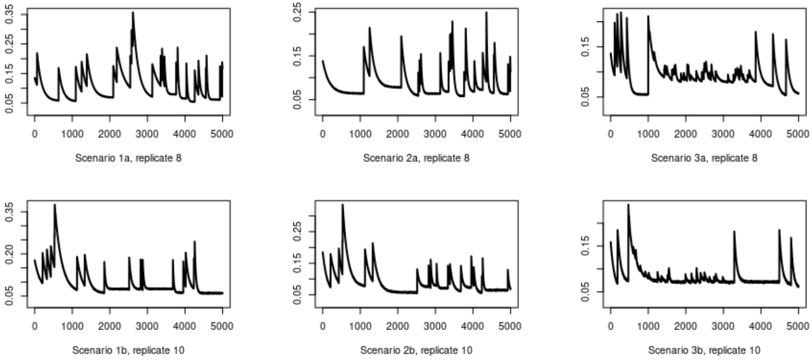

Figure: Examples of the simulated time series from six scenarios.

Simulation study - result

Table: Number of simulation replicates with true positive rates $\geq 80\%, 90\%$

	True positive		**True positive**	(+/- 10)
	$\geq 90\%$	$\geq 80\%$	$\geq 90\%$	$\geq 80\%$
S1a	142 (200)	174 (200)	154 (200)	192 (200)
S2a	119 (200)	181 (200)	146 (200)	190 (200)
S3a (fine)	28 (100)	68 (100)	40 (100)	83 (100)
S3a (large)	81 (100)	95 (100)	83 (100)	97 (100)
S1b	133 (200)	174 (200)	153 (200)	196 (200)
S2b	116 (200)	178 (200)	143 (200)	191 (200)
S3b (fine)	46 (100)	77 (100)	50 (100)	85 (100)
S3b (large)	77 (100)	95 (100)	80 (100)	95 (100)

Simulation study - result

Table: Summary statistics of the distance measure [Shi *et. al.*(2021)] between two sets of changepoints, the root mean squared errors (RMSE) of the estimated decay parameter γ.

	Distance			**RMSE γ**		
	10%	median	90%	10%	median	90%
S1a	0	0.0015	1.2325	0.0001	0.0016	0.0936
S2a	0	0.0227	2.1863	0.0002	0.0124	0.0926
S3a (fine)	0.0053	1.1093	5.0979	0.0147	0.0844	0.1341
S3a (large)	0	1.0019	4.0000	0.0006	0.0063	0.0401
S1b	0	0.0053	1.2023	0.0002	0.0023	0.0980
S2b	0	0.0343	2.1302	0.0003	0.0147	0.0934
S3b (fine)	0.0128	3.0703	23.0000	0.0110	0.0713	0.1379
S3b (large)	0	1.0000	2.0912	0.0006	0.0032	0.0174

Application to NEON soil moisture time series

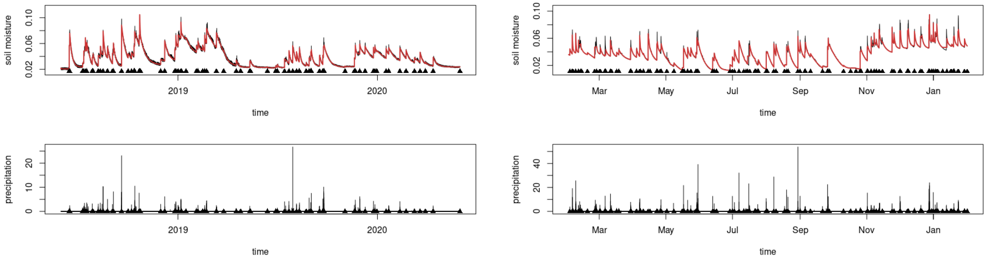

Figure: (Top panels) The identified changepoints (black triangles) and the modelled time series (red curves). (Bottom panels) The rainfall time series with overlaid changepoints (black triangles).

Relax the assumptions

Not all segments in the soil moisture time series follow the exponential decay assumption. For example, the soil water content may fluctuate around certain level for weeks.

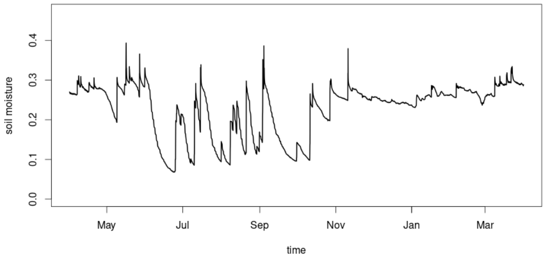

Figure: A example of soil moisture time series with a winter period that does not follow the exponential decay assumption.

Changepoint detection using multiple models

- We would like to have a changepoint detection procedure that can **choose between multiple models** to describe different patterns show in different segments, e.g. an exponential decay model for the well drained period, and a simple mean model for the saturated periods.
- A changepoint model that can detect changes in both the types of model and the model parameters.
- A method based on the **Bayesian changepoint detection** detailed in [Fearnhead & Liu(2007)] has been investigated.

Changepoint detection using multiple models

- Denote C_t as the last changepoint prior to time t and M_t as the model index for the segments terminating at time t. The goal is to find $p(C_t, M_t | Y_{1:n})$.
- The joint distribution of the observation Y_t and the latent states C_t, M_t is

$$\prod_{t=1}^{n} \{f(Y_t | C_t, M_t, \theta_t) \times p(C_t, M_t | C_{t-1}, M_{t-1}, \pi)\} \times priors$$

  where θ_t is segment specific parameter, and $p(C_t, M_t | C_{t-1}, M_{t-1}, \pi)$ is the transition distribution with parameter and π.
- Inference uses the **sequential Monte Carlo** method following [Fearnhead & Liu(2007)] and [Fearnhead & Liu(2011)].

Changepoint detection using multiple models

- Specifically, the **forward filtering** recurssions can be written as

$$p(C_{t+1} = j | Y_{1:(t+1)}) \propto \begin{cases} w_{t+1}^{(s)} p(C_{t+1} = s | C_t = s) p(C_t = s | Y_{1:t}) \ , \ j = s \\ w_{t+1}^{(t)} \sum_{s=1}^{t-1} p(C_{t+1} = t | C_t = s) p(C_t = s | Y_{1:t}) \ , \ j = t \end{cases}$$

where $w_t^{(j)} = f(Y_{t+1} | C_{t+1} = j, Y_{1:t})$ is calculated as

$$\frac{\sum_m f(Y_{(j+1):(t+1)} | C_{t+1} = j, M_{t+1} = m) p(M_{t+1} = m)}{\sum_m f(Y_{(j+1):t} | C_t = j, M_t = m) p(M_t = m)}$$

- The **backward simulation** relies on the forward filtering distribution at each changepoint location.

Challenges on parameter estimation

- Static parameters, or the global parameters, can be estimated using particle MCMC as described in [Whiteley & Andrieu(2009)].
- Segment specific parameters, such as θ_t, are much harder to estimate when they cannot be integrated out of the log-likelihood.

We are able to estimate the segment specific parameters given the changepoints. However, the changepoint detection itself relies on the choice of intial parameters. This is a problem we are still working on.

References

- Fearnhead, P., Liu, Z., 2007. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*. 69 (4), 589–605.
- Fearnhead, P., Liu, Z., 2011. Efficient Bayesian analysis of multiple changepoint models with dependence across segments. *Statistics and Computing*. 21, 217–229.
- Jewell, S., Witten, D., 2018. Exact spike train inference via $\mathcal{L}_0$ optimization. *Annals of Applied Statistics*, 12(4), 2457–2482.
- Jewell, S., Hocking, T. D., Fearnhead, P., Witten, D., 2020. Fast nonconvex deconvolution of calcium imaging data. *Biostatistics*, 21(4), 709–726.
- Killick, R., Fearnhead, P., Eckley, I. A., 2012. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500), 1590–1598.

References

- McColl, K. A., Wang, W., Peng, Bin., Akbar, R., Gianott, D. J. S., Lu, H., Pan, M., Entekhabi, D., 2017. Global characterization of surface soil moisture drydowns. *Geographical Research Letters*, 44, 3682–3690.
- Ruscica, R., Polcher, J., Salvia, M., Sorensson, A., Piles, M., Jobbágy, E., Karszenbaum, H., 2020. Spatio-temporal soil drying in southeastern South America: the importance of effective sampling frequency and observational errors on drydown time scale estimates. *International Journal of Remote Sensing*, 41:20, 7958-7992, DOI: 10.1080/01431161.2020.1767825
- Shi, X., Gallagher, C., Lund, R., Killick, R., 2021. A comparison of single and multiple changepoint techniques for time series data. *arXiv:2101.01960*.
- Whiteley, R. C., Andrieu, M. R., 2009. Particle Markov Chain Monte Carlo for Multiple Change-point Problems. *Department of Mathematics, Bristol University, Bristol, UK, Technical Report*. 911, 26.

Appendix A: more on the simulation study

The distance measure used in the simulation study was developed by [Shi *et. al.*(2021)]. It quantifies the differences between the two sets of changepoints, e.g. the true changepoint set $\mathcal{C}_0 = \{\tau_1, \cdots, \tau_m\}$ and the estimated changepoint set $\mathcal{C}_1 = \{\eta_1, \cdots, \eta_k\}$, and is defined as

$$d(\mathcal{C}_0, \mathcal{C}_1) = |m - k| + \min\{\mathcal{A}(\mathcal{C}_0, \mathcal{C}_1)\} \ ,$$

where m and k are the number of changepoints in each set, and

$$\mathcal{A}(\mathcal{C}_0, \mathcal{C}_1) = \sum_{i=1}^{m}\sum_{j=1}^{k} c_{ij} I_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{(\tau_i - \eta_j)}{N} I_{ij},$$

is the overall cost of assigning η_j to τ_i, $j = 1, \cdots, k$, $i = 1, \cdots, m$. In particular, $I_{ij} = 1$ if η_j is assigned to τ_i and $I_{ij} = 0$ otherwise, following a linear assignment problem.

Appendix B: more on the application

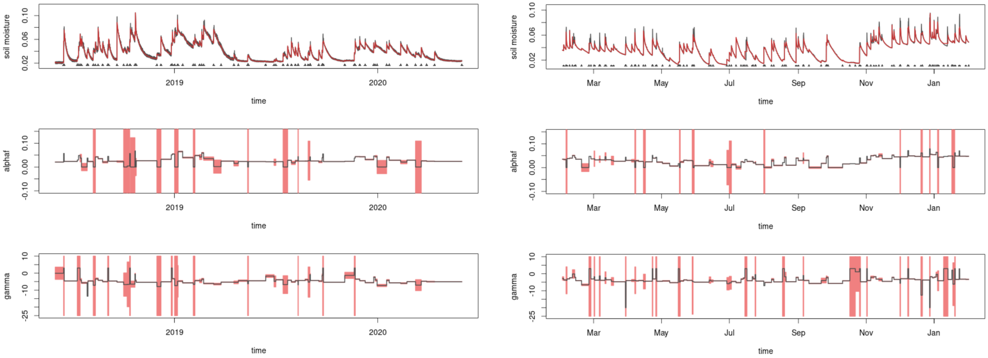

Figure: The estimated parameters of the exponential model plotted over time along with their uncertainties (from nls estimation). The middle panels show α_f and the bottom panels show γ.

Appendix C: simple illustration of the 2nd model

An experiment on simulated soil moisture time series. The exponential decay rate was fixed for all segments and was estimated using particle MCMC.

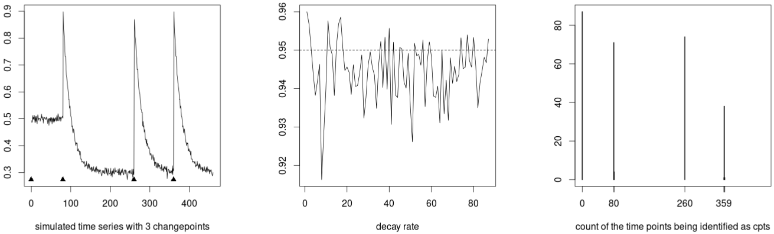

Figure: The simulated time series (left), the estimated decay rate (middle), and the count of the etimated changepoints from the accepted proposals of the last 100 iterations.